Author:
Alexander, Keith

Title:
Projecting proteins and random walks
knotting in open curves via virtual knots

Projecting proteins and random walks: knotting in open curves via virtual knots

---

Keith Alexander

H. H. Wills Physics Laboratory
University of Bristol

A dissertation submitted to the University of Bristol in accordance with the requirements for award of the degree of Ph.D. in the Faculty of Science

August, 2018

Word count: 39,001

Abstract

In this thesis we develop a new method of knot recognition for open curves based on taking many projections and identifying them as virtual knots, an extended class of knotted objects which exist 'in-between' classical knot types. We call this method virtual closure. We explore how virtual closure differs from a method we call sphere closure which involves joining the ends of the curve to many far away points, finding that virtual closure is more sensitive to knotting and provides more complex and detailed conformational information. An important distinction we find is between curves which present a single dominant knot type across closures, which we call strongly knotted, and more ambiguous curves which are knotted but with many different knots depending on the closure chosen, which we call weakly knotted.

We perform a knotting survey of all proteins in the Protein Data Bank using virtual closure. Compared to previous sphere closure surveys, we find 25% more knotted proteins. Of all the knotted proteins, 40% are found to be weakly knotted under virtual closure, many more than under sphere closure, hinting that the knotting in proteins is more ambiguous than was previously thought.

We then investigate the knotting of random walks, finding that weak knotting is very rare in unconfined walks, but increasingly common in isotropically confined walks both on and off-lattice. We determine that weak knotting is essentially length independent, instead depending only on the degree of confinement - the ratio of average radius of gyration of unconfined walks to confined walks of the same length. The greater the degree of confinement, the more likely that knotting is weak. We reduce the number of confined dimensions, moving from walks in the sphere, to the tube and then to slits, finding that overall knotting and weak knotting become less common.

Acknowledgements

My first thanks have to go to my supervisor Mark Dennis for the chance to do the PhD and for all the support and ideas throughout. In particular, I would have achieved far less without his encouragement to push myself in all areas. While it wasn't always painless, worthwhile things often aren't and Mark always had my back.

I owe a particular debt to Sandy Taylor for being so helpful throughout. He put up with my inane coding questions, helped me analyse and understand my results and kept me honest whenever I was reluctant to take the hard road always with good humour and eloquence. I learned much through his example.

Thanks too to all the SPOCK team in Bristol, in particular Dave Foster, Ben Bode, Danica Sugic, Elena Boniolo, Lauren Scanlon, Emma Creasey and honourary SPOCK member Teuntje Tijssen. You have all been such great company and support for the last few years. Thanks to Simon Hanna also for supervision particularly in Mark's absence. While it's a shame the polymers work didn't quite work out, I still learned a lot doing it.

A special thanks to Alex Houston, gone but not forgotten. His friendship, late night talks and cups of tea kept my spirits up and he is a miss now he's moved on to pastures new in Warwick.

While much has changed in my time in Bristol, the biggest change has been meeting my wonderful fiancée Polly Foster. She has made the PhD so much easier with her love and support and I look forward to spending the rest of my life with her.

Lastly, thanks to Mum and Dad. While they might not know much knot theory, they know about the other important things in life and have always believed in me even when I have not. Thanks for guiding me towards being the person I am today.

Author's declaration

I declare that the work in this dissertation was carried out in accordance with the requirements of the University's *Regulations and Code of Practice for Research Degree Programmes* and that it has not been submitted for any other academic award. Except where indicated by specific reference in the text, the work is the candidate's own work. Work done in collaboration with, or with the assistance of, others, is indicated as such. Any views expressed in the dissertation are those of the author.

SIGNED: ..... DATE:.....

Keith Alexander

Contents

List of Figures

1 Introduction

From a very young age knots have been a part of our lives. At the time, tying one's shoelace was perhaps the most difficult task of coordination and manipulation ever attempted. The invention of knots marks a very significant point in the early history of human development also, allowing for the innovation of other technologies that depend on them. While we may no longer have to secure an axe head to a shaft with string, we are familiar with the use of knots in applications from seafaring to knitting. In this way, many of the physical effects of knots are well known and intuitive for us, providing stability and additional functionality to otherwise plain pieces of rope, and confounding music listeners with the dense, compact tangle of headphones.

Given how readily any string or cord will self-entangle, it is not a surprise that knots are an unavoidable feature in many string-like systems, with the world of *polymers* providing rich examples. Polymers are long molecules made from smaller building blocks called *monomers*, and they can have incredibly diverse properties depending on the monomer or combination of monomers which make them up and the environmental conditions they are in [1]. They form many useful materials such as rubber and plastics as well as myriad biomolecules vital to life such as DNA, RNA, polysaccharides, cellulose and proteins.

Given the importance of polymers, much research has been directed to understanding all aspects of their physics, both of the bulk materials they make up and of isolated molecules. This has included work related to the entangling of polymers [2, 3, 4, 5, 6, 7, 8]. It was conjectured in the early 1960s that knotting would be inevitable in all sufficiently long and flexible polymers [9, 10]. In the late 1980s this was proved mathematically for some models of polymers [11, 12] and there has been intense interest since trying to determine what this knotting

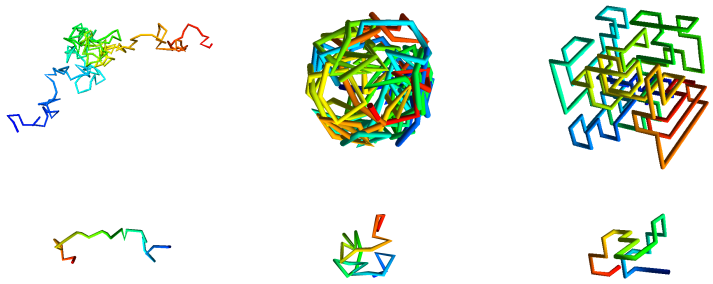


Figure 1.1: Examples of the open random walks we will be investigating later.

looks like, how it arises in different systems, what affects knotting and how knotting affects the properties of polymers [3].

Often in trying to understand flexible polymers it is useful and easier to model the polymer as a *random walk* instead of a more detailed physical model [13]. Random walks in this sense are essentially random curves in three-dimensional space, or 3-space. They can be tailored to include stiffness, solvent quality, to exclude volume, create rings, lie on a lattice or continuous space, exist in confinement or all manner of other properties which can help more accurately mimic real, specific polymers or ease computation. Some examples of random walks are given in Fig 1.1. It is for self-avoiding random walks on the cubic lattice that the inevitability of knotting with increasing length was first proved [11, 12] and they continue to be a fruitful source of insight into the knotting of polymers.

Not all polymers can easily be modelled by random walks however, with *proteins* being a particularly important example. While some disordered proteins do behave like free, flexible polymers [14, 15, 16], many other essential proteins are highly structured [17] like those in Fig 1.2. Upon being created in the cell, proteins undergo a process called *folding*, during which they orient their backbone into a very specific shape, often securing this with additional bonds between non-neighbouring monomers. This shape allows proteins to perform their many varied functions and so there is great interest in understanding how they fold and what structures they form [18]

Initially it was thought that proteins could not form knots [20] since knot formation would require a more complex folding pathway and increase the time and energy taken to produce the protein compared to a protein not containing a knot. Hence, if a protein could be created to perform the same function without a knot, this should be evolutionarily preferred. Nevertheless, knotted

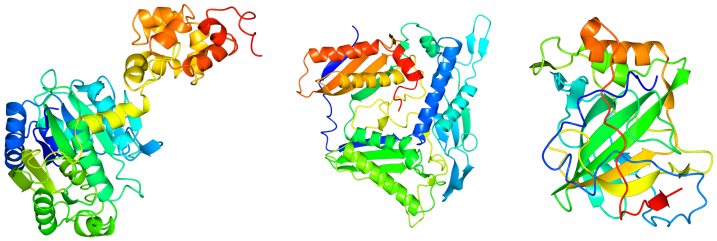


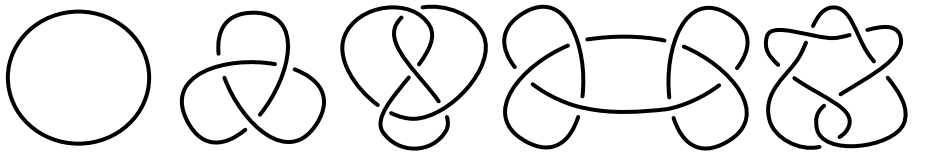
Figure 1.2: Examples of the proteins we will be investigating later. These diagrams were created with CCP4mg [19].

proteins have been found [21], raising questions about how they fold and what properties the knots confer which cannot be obtained otherwise. These are still contentious matters with a full understand not yet reached. What does seem to be clear is that in comparison to flexible polymers and random walks, knotting in proteins is rare [22, 23]. In order to understand this knotting, it is crucial that we can distinguish different knotted structures in as much detail as possible.

In order to understand the knotting of any space curve, let alone proteins and other polymers, we need a firm, mathematical understanding of what a knot actually is. The mathematics of *knot theory* forms a branch of *topology*, the study of the shapes of objects and how they may be deformed continuously. In order to study knots topologically, a space curve must be closed, forming a single loop. Knot theory then is the study of these loops and which loops can be deformed to look like each other without cutting or glueing [24].

There are infinitely many different topologically distinct knots that one may tie in a *closed* loop. These range from the trivial planar circle, called the *unknot*, to well known practical knots such as the *trefoil* knot, the *figure-eight* knot, the *reef* knot and the *stevedore* knot and beyond to arbitrarily complex tangles. Examples are given in Fig 1.3. In this figure we represent the true three-dimensional curve forming the knots with a two-dimensional projected diagram called a *knot diagram*. The only 3D information remaining is where strands of the diagram cross each other, where it is important to note which strand passes over which. For a closed curve, all that is needed to determine the type of knot formed is a single projected knot diagram.

This is extremely useful when dealing with ring polymers which are naturally closed and cannot cross themselves without undergoing some sort of physical process. However, many interesting polymers such as proteins are not rings but linear, *open curves*. In the true, mathematical sense, these cannot contain knots.



Unknot	Trefoil	Figure-eight	Reef knot	Stevedore knot
0_1	3_1	4_1	$3_1\#3_1$	6_1

Figure 1.3: Examples of topologically distinct knots. None of these may be deformed to look like the others without cutting and glueing. The upper text gives the common English name for each knot, and the lower text the mathematical label. The first number in each label gives the number of crossings in the knot, and the subscript label is an arbitrary index to distinguish between distinct knots with the same number of crossings. As the reef knot is made from tying two trefoil knots together, its label reflects this. We will discuss these labels more in Chapter 2.

Topologically, all open curves are equivalent as any tangle can eventually be unthreaded and the curve reduced to a straight line. In more mathematical terms, knots themselves are often not dealt with directly, but rather the space outside of the knot, known as the *knot complement*. This is the space obtained from 'drilling out' the knot from 3-space and the knot complements of topologically equivalent knots are also topologically equivalent. The complement of all open curves are topologically equivalent and strikingly the 'hole' left by the open curve can always be deformed to a point without changing this topology. This is not true of true closed knots which cannot be removed by deformation and is another way of highlighting the difference between open and closed space curves.

Nevertheless, there is an intuitive sense in which we say that our shoelaces are knotted, despite the fact we do not glue the ends together. In such cases we can instinctively point to where the knot is and recognise that our untied shoelaces are different in a significant way from their tied conformations.

While open curves can never be topologically knotted, they can bear a close geometrical resemblance to knotted closed curves. Several methods have been proposed to capture this geometrical resemblance, all of which at some point involve joining the open ends together to create a closed curve which can be analysed [25, 26, 27, 28]. Naturally, this may raise concerns that the resulting type of knot on closure may not be accurate if the closure threads or unthreads parts of the open curve. The method which has become standard to avoid this problem involves taking multiple closures of the curve and looking at the

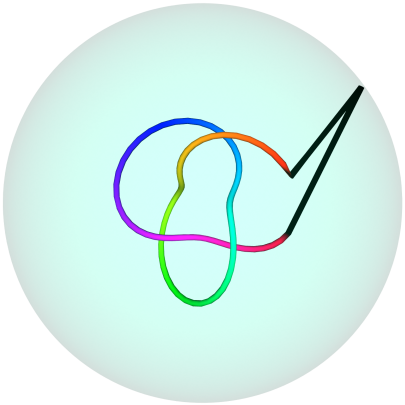


Figure 1.4: A simple open curve being closed with straight lines to a point on a sphere centred on the curve. Such a closure turns the topologically trivial open curve into a closed trefoil knot.

spectrum of knots obtained [26]. Fig 1.4 shows how a simple curve can be closed to a point on a sphere surrounding the curve. If 100 different closure points are chosen uniformly on this sphere, this is usually sufficient to capture the knotting of the curve. In cases like the average shoelace where the knot is tight and the ends are extended from the knotted area, a single type of knot will appear in most closures and this can simply be taken as representative of the knotting of the open curve. This has worked well for many polymers of interest so far, but there are conformations where the ends of the polymer are not so extended from the tangled area. In these cases there often is not a single type of knot which appears to dominate over closures and it is difficult to say that there is a particular type of knot tied in the open curve [26, 28].

One might wonder if there is a natural way to capture these ambiguous, mixed knotted configurations. One may also wonder if there is a way to recognise knots in open curves that doesn't involve adding a physical closure between the ends of the curve. There is a more abstract mathematical object that can help us do this, called a *virtual knot* [29]. Virtual knots cannot be tied in a piece of string like an ordinary knot, but they can help us categorise the knotting of projections of open curves. It will suffice for now to consider virtual knots as existing 'in-between' other, ordinary closed curve knots like those in Fig 1.3, which we now call *classical knots*.

For example, say we have an open curve in the conformation shown in Fig 1.5. From the viewing direction presented, we can take a projection to get an open knot diagram. Comparing with the classical knots we drew in Fig 1.3,

it is difficult to say whether the unknot or the trefoil knot best represents the knotting of the curve when viewed from this direction. It appears more knotted than the unknot, but less knotted than the trefoil knot. There is a virtual knot, which we call the *virtual trefoil*, which we can use to classify this ambiguous, in-between configuration. As can be seen in the figure, we close the projected diagram and add a circle where we have crossed an existing strand to note that this crossing was not initially there. Note this is not a physical closure of the curve in 3-space, we are just using virtual knots to distinguish different projected diagrams. We are not analysing anything that was not present in the curve to begin with. A complete virtual analysis of an open curve would involve taking many projections, just as many closures were taken previously, in order to fully capture the knotting of the curve.

This approach to knot recognition was originally suggested by Dr. Alexander Taylor and forms the basis for the knotting analysis performed in this thesis. As we will see, using virtual knots gives us a method which is more sensitive to knotting, detecting knots in less tangled curves, and more sensitive to ambiguities in which knot is tied than methods which involve closing to the sphere.

We stress here again that open curves are all topologically equivalent and trivial. Attempting to capture the geometrical resemblance to knotted closed curves with a single closure can lead to different answers depending on how the curve is closed. Taking multiple closures, either physically to the sphere or virtually on projection, from all directions gives us an ensemble of closed knots which can be topologically distinct. By analysing this ensemble statistically we can obtain a reliable, objective measure of the geometrical similarity between open curves and closed curves. When we talk about the knotting of an open curve, we are referring to this ensemble.

1.1 Proteins background

Having given a flavour of the themes of the thesis, we now give some more background on the key open curve systems we will be examining, starting with proteins.

1.1.1 Structure of proteins

Proteins are complex biomolecules which perform a huge variety of functions in every living organism. They are made from at least one chain of *amino*

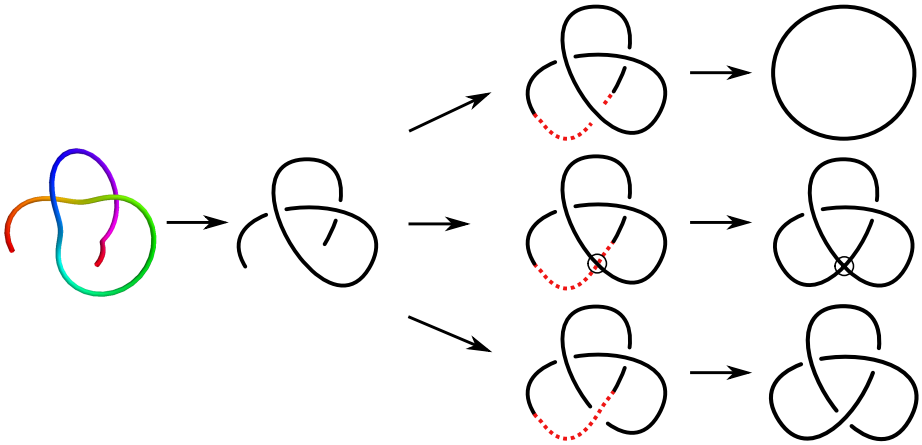


Figure 1.5: An open curve, left, with its projected diagram, middle left. If we close the ends of this diagram with an under-crossing (top), we obtain an unknot. If instead we choose an over-crossing (bottom), we obtain a trefoil knot. The middle path shows a virtual closure resulting in a 'virtual trefoil knot'. This is less knotted than a trefoil but more knotted than an unknot, capturing the nature of the original projection.

acids, and it is not uncommon for many chains to bond together to create the final protein. These can range in size from the enormous Titin, which contains up to 33,000 amino acids and can be a micrometer long [30] to the tiny Trp-Cage, which is only 20 amino acids long [31]. The final structure and function of each protein is dependent on which amino acids appear and where along each chain [18]. Different amino acids have different functional groups attached, presenting different internal bonding opportunities and final active sites. Determining and predicting protein structure is a large effort in modern biology and remains a difficult problem despite the ever increasing computational resources available. The process of protein folding by which unstructured amino acid chains take on the final form of the protein takes place in a highly crowded cellular environment and while the basic physics is well understood, the complexity of this environment makes simulation and prediction incredibly challenging.

Fig 1.6 shows the basic structure of an amino acid, the key components of which are the *amino group* at one end, the *carboxyl group* on the other, and the side chain between them. The R group in the side chain is different for each amino acid and is the source of their diversity and function. It is attached to what is called the *alpha-carbon*, or C_α. While there are many possible amino acids, there are only 22 which are known to be genetically coded for, called the

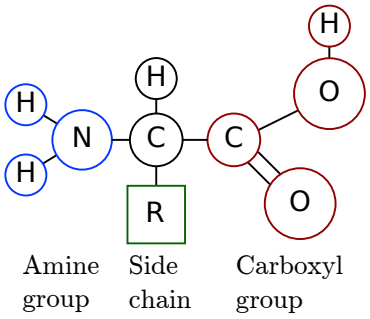


Figure 1.6: The generic structure of an amino acid. R refers to the functional group of the amino acid. All other letters are chemical element symbols.

proteinogenic amino acids. These are what all known proteins created by living organisms consist of [18].

The sequence of bases in a section of DNA encodes the sequence of amino acids in a protein. *Translation* is the process of turning this code into an amino acid chain and it occurs at an organelle called a *ribosome*. A condensation reaction joins each amino acid together with a covalent bond, known as a *peptide bond* or *amide bond*, as shown in Fig 1.7.

The full sequence of amino acids in a protein is known as its *primary structure*. This structure gives proteins a natural orientation, as one end terminates in an amine group, the *N-terminus*, and the other in a carboxyl group, the *C-terminus*. The protein is translated from the N-terminus to the C-terminus, and so the N-terminus end is free to begin folding into more complex three-dimensional structures while the rest of the chain is still being formed.

The first structures to form during or after translation are ordered on a relatively local scale and typically characterised by their hydrogen bonding. The most common motifs in this *secondary structure* of the protein are *alpha helices* and *beta pleated sheets*, although other structures have also been identified. Alpha helices are spiral structures, where a single turn is composed of roughly 3.6 amino acids. Hydrogen bonds between amino acids separated by a turn of the helix stabilise the structure. Beta sheets are formed from elongated arrangements of amino acids known as *beta strands*. These strands can be folded back on themselves and hydrogen bonded to each other to create beta sheets.

Once these structures have formed, they can then be arranged into the *tertiary structure* of the protein. If a protein consists of only one chain, this is the final stage of folding. Much of this stage is determined by hydrophobic

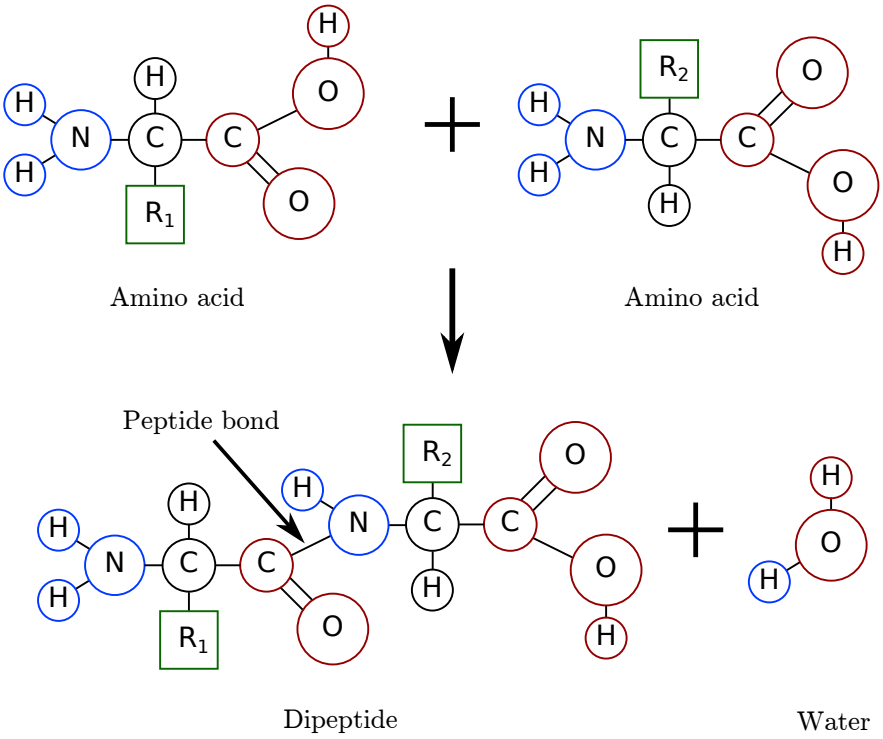


Figure 1.7: The reaction of peptide bonding two amino acids together.

interactions, where the hydrophobic patches on the alpha helices and beta sheets arrange to face each other and away from the aqueous environment of the cell. This structure is stabilised by further hydrogen bonding and in some cases by covalent bonds called *disulphide bridges*, or *cysteine bridges*, so-called because they form between two cysteine amino acids. If the protein consists of more than one chain, then these arrange themselves after taking on their tertiary structure to form the final *quaternary structure* of the protein. These levels of protein structure are shown in Fig 1.8.

The structure of an actual protein, PDB ID 4COQ, is shown in Fig 1.9. As can be seen, the protein can be represented in many different ways, from positions of every atom in a) to the backbone outline in c). The *ribbon diagram* shown in Fig 1.9 b) is a particularly common and useful representation, providing a good compromise between an easy to see backbone shape and additional bonding information in the important alpha helices and beta sheets.

As the energy landscape being explored during the process of protein folding contains many local minima giving incorrect structures, folding is susceptible to mistakes or misfolding. There are mechanisms present, such as *chaperone* proteins, to reduce the rate of misfolding and to help misfolded proteins back-

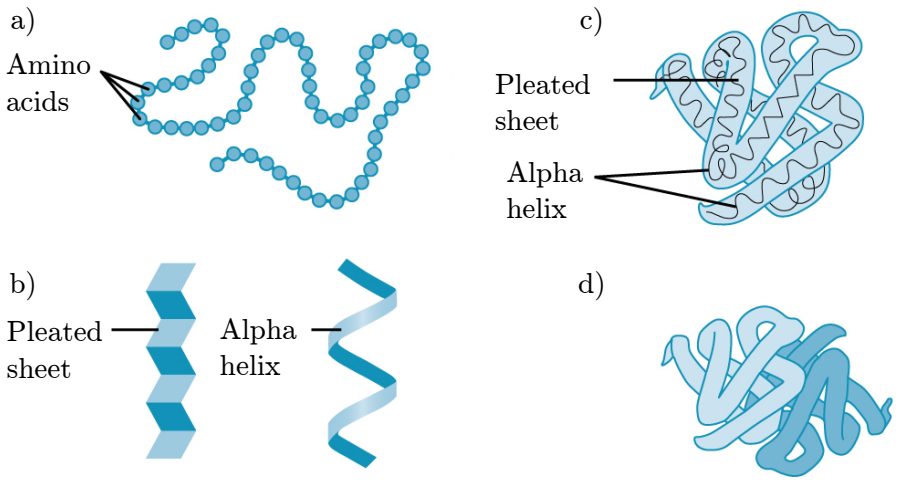


Figure 1.8: Different levels of protein structure. a) primary structure, b) secondary structure, c) tertiary structure and d) quaternary structure. Figure adapted from [32].

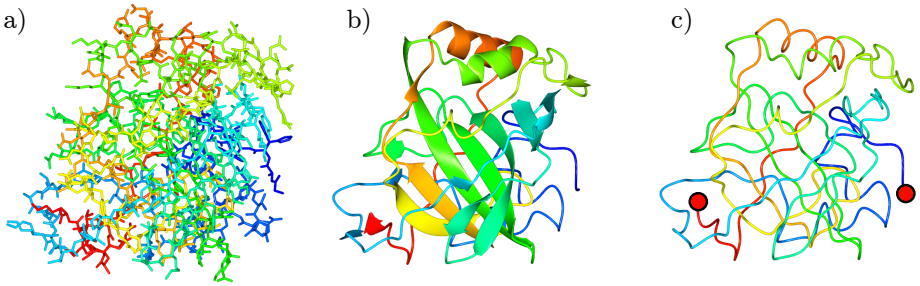


Figure 1.9: The same protein represented in different ways. a) shows a stick diagram of the protein where a different atom is placed at each vertex. b) shows a ribbon diagram, where alpha helices are represented by the thickened coils and beta sheets by arrows. c) just the open curve traced by the backbone of the protein. The end-points are picked out by the red circles.

track to a point where they can try again. Some proteins can fold correctly without the aid of chaperones, but this is not true for all proteins. The effects of misfolding can range from the protein taking a little longer to achieve the right conformation, to not performing its function correctly, to aggregating with other misfolded proteins to cause diseases such as Alzheimer's disease or Parkinson's disease [33].

1.1.2 Knotted proteins

While protein structures have been available for around 60 years now, it is only in the last 20 years that researchers have begun to look at knotted structures

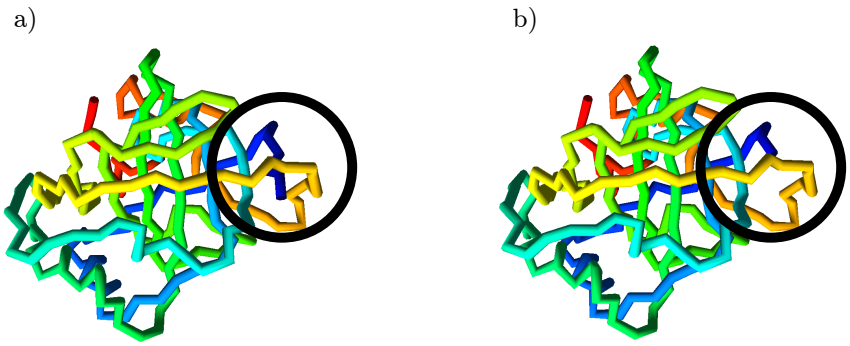


Figure 1.10: One of the proteins considered by Mansfield for knotting [20]. a) shows the complete structure which shows detectable knotting. b) shows the same protein but with 3 residues removed from the highlighted end. This simple procedure removes the detectable knotting of the protein.

fruitfully. There is a remark made in parentheses about a knotted protein, carbonic anhydrase C, in 1977 from Richardson [34]. This is followed by another remark in parentheses, 'Presumably a knot is not impossible if the piece to be tucked through the loop is very close a chain end.' Other than this tentative example though, the consensus was that knots were not possible in proteins.

The first systematic search for knotted protein backbones in the *Protein Data Bank* (PDB), the online repository for all resolved protein structures, in 1994 by Mansfield [20] found that, of the approximately 400 structures deposited at the time, only three showed detectable knotting. Mansfield looked closer at the three knotted examples and concluded that in each case, the knot was formed, 'by tucking several residues at one end through a wide loop passing around the exterior of the molecule.' He did not consider these as truly knotted as the removal of a few residues can remove the knot, as in Fig 1.10. The consensus at this point was that the formation of a loop and the subsequent threading to form a knot was too difficult a folding procedure to occur reliably [35, 36].

This consensus was challenged in 2000 with Taylor's discovery of a 'deeply knotted' protein [21], deeply knotted here meaning that knot can be untied only after removing 70 amino acids from the C-terminus, or 245 amino acids from the N-terminus. In addition to this protein being indisputably knotted, where the knot could not be explained away as a chance folding of the very end, it contained a figure-eight knot. This was the first time anything more complicated than a trefoil had been seen and it showed that protein structure could indeed be topologically complex, starting a more concerted research effort focussing on this.

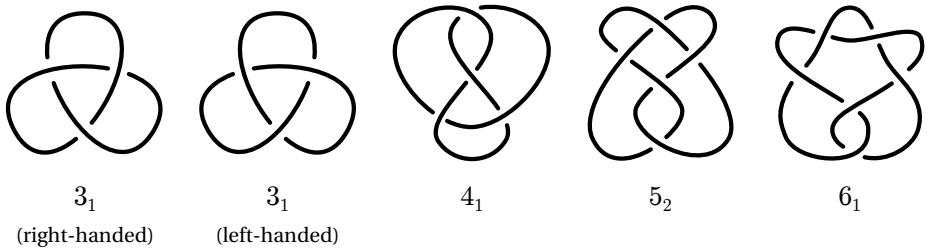


Figure 1.11: All the different types of knots currently observed in proteins with their usual mathematical labels. The trefoil and its mirror image are distinct knots and have both been found.

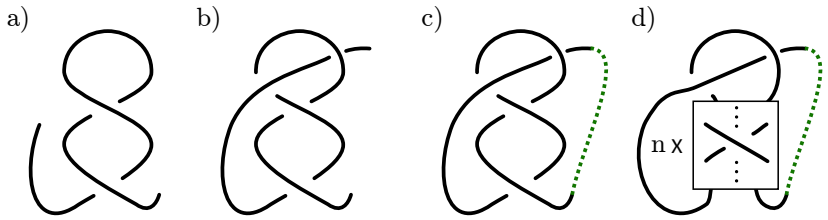


Figure 1.12: A potential generic folding pathway for knotted proteins. a) A twisted loop. b) A threading of this loop. c) The obvious closure of this curve to give, in this case, the knot 4_1. d) The general form of twist knots. The box contains arbitrarily many twists with each number producing a different knot. One twist produces the trefoil, two the figure-eight, three the knot 5_2, and four the stevedores knot 6_1, accounting for all protein knots.

Since Taylor's finding, protein backbones have been found to form more complex knots, the complete list given in Fig 1.11. Various protein knotting databases have since been developed to catalogue this, including KNOTS [37] and pKNOT [38] although it appears that pKNOT is no longer accessible as of 22nd June 2018. The most successful database currently is *KnotProt* [23, 39], which maintains an up-to-date record of knots found in the PDB. KnotProt also catalogues *slipknotted* proteins, where the complete structure is unknotted, but removing residues reveals the knotted structure.

There are still many unanswered questions about knotting in proteins, including how they fold. All knots currently observed in proteins are *twist knots*, which can be formed by making a loop and twisting the strands around each other a number of times before embracing the loop with one of the loose ends, as illustrated in Fig 1.12. This suggests an obvious method of protein knot folding [40], where all topology enters the chain through a single threading move.

The details of the folding are still unclear however. Jackson's group did

pioneering work in understanding this, first showing that knot formation can be spontaneous [41], in that it can occur without misfolding with no assistance from chaperone molecules, but that the knotting rate is significantly increased in the presence of chaperones. They later showed that the threading of the loop to form the knot in a particular protein always occurs from the same terminus [42]. Blocking this path to folding by binding a large molecule to the threading terminus resulted in the protein failing to fold. Simulation of these processes, as with all protein folding, remains difficult however.

Another key open question regards the function of knots in proteins. Knotted proteins are slow to fold, given the relative complexity of their folding pathways, and so are expensive for organisms to produce. If an unknotted alternative could be produced, this would likely be highly evolutionarily favourable. While knotting of proteins is rare, its presence suggests that indeed there must be a functional benefit to knotting. Denatured proteins retain their knotted backbone [43], suggesting a stabilising effect. This is borne out in other experiments suggesting a thermal and chemical stability to knotted proteins [44, 45, 46, 47]. Further, most knotted proteins are enzymes [23, 48]. It has been suggested that the rigidity and geometry given by the knotted backbone serves to present a favourable active site configuration [49]. Much of this work is still quite speculative however. A good summary is provided by Dabrowski-Tumanski and Sulkowska [48].

More complicated features of protein topology have also been explored by including not just the peptide bonds along the backbone chain but also disulphide bridges between more distant chain sections. This branched structure can be understood as a three-valent spatial graph. By considering cycles of this graph, true topological knots can be found, but there is also interest in looking for new types of structures such as *links* [50, 51, 52] and *lassos* [53, 54, 55]. Links appear dramatically in the viral capsid of bacteriophage HK97, where 72 separate proteins link in a chainmail fashion to form a sphere [56]. Other, smaller examples of links can be found in other PDB structures [50, 44], with LinkProt providing a useful database for these [51] Lassos are a new 'topological' object consisting of a closed ring which is pierced by one or both of the loose ends of the protein. LassoProt provides a database of these structures [55] and as is common with all topological features of proteins, studies are ongoing as to the function of these motifs.

1.2 Random walks background

The other systems we will be examining closely are random walks. A random walk is a path created by taking random steps in some mathematical space. For these walks to be knotted they must be in a three-dimensional space such as $\mathbb{S}^3$, $\mathbb{R}^3$ or $\mathbb{Z}^3$. Each step may in general be of any length, but the walk may not intersect itself for knotting to exist. While in this thesis we have a measure for knotting in open curves, walks can be made to close and studied as regular classical knots. Closed random walks are often called *random polygons*.

Random walks have long been of interest as models for various physical systems and processes [13, 3, 6]. Particular interest in the knot sustaining walks we look at has come from the polymer community. It is difficult to study experimentally the microscopic motions and conformations of polymers, and full theoretical treatments are also very challenging. To this end, random walks have been used as a model for polymers, with different polymer properties able to be considered by varying random walk parameters [13]. In some of the literature, polymers and random walks are almost used interchangeably, so close is the connection.

Examples of random walks include *ideal chains* which are walks in $\mathbb{R}^3$ where each step is taken in a uniform random direction and each step is of equal length. Walks with equal length steps are often called *equilateral walks*. Ideal chains are often used to model polymers under *theta conditions*, where the polymer is neither collapsed and globular, or in the coiled state where it is expanded and more disperse [13]. In this case, each step of the walk represents several monomers in the polymer, at least as long as the *persistence length* of the polymer and so the direction of each step of the walk is uncorrelated with the last.

DNA is a polymer that is often modelled as an ideal chain. One aspect that may concern a reader is that, in an actual polymer system that is evolving dynamically over time, strands cannot pass through themselves and so the statistics of ideal chain conformations may not accurately tell us the statistics of polymer chain conformations, especially if those chains are closed. DNA is an interesting example in this respect as there exist enzymes, called *topoisomerases*, which can cut a strand of DNA, allow another strand to pass through, and then rejoin the cut strand. In a topoisomerase rich environment, DNA behaves more like a *phantom* polymer, where the strands are allowed to cross each other freely,

making the ideal chain model more valid [57, 58].

The step length of walks in $\mathbb{R}^3$ can also be drawn from some distribution, rather than being uniform. A common choice is a Gaussian distribution, with walks being called *Gaussian random walks*. The Gaussian walk can be obtained from an equilateral walk, where each step in the Gaussian walk corresponds to several steps of an ideal walk. As the distance between end-points in an ideal walk is Gaussian distributed, a Gaussian walk is thus obtained [1].

Another important class of random walks are walks on the cubic lattice, which is a particular embedding of $\mathbb{Z}^3$. Enforcing the condition that these walks cannot cross themselves creates a naturally *self-avoiding* walk. Unless extra conditions are applied, walks in $\mathbb{R}^3$ as described have the potential to intersect themselves. While this happens with probability approaching zero for finite length walks and so isn't a concern from a knotting perspective, it is often desirable to imbue walks with a self-avoiding property when modelling physical systems which are naturally self-avoiding. This can be done in many ways, such as enforcing a tubular region around each section of the walk which cannot be intersected by the rest of the walk [59]. While the small length scale properties of lattice walks and self-avoiding walks in $\mathbb{R}^3$ are obviously different, at longer lengths they behave very similarly [1].

The size of random walks depends intimately on their self-avoidance or lack thereof. The *radius of gyration* of a random walk is defined as:

$$R_g^2 = \frac{1}{N}\sum_{k=1}^{N}(\mathbf{r}_k - \mathbf{r}_{mean})^2 \tag{1.1}$$

where r_k is the position of the k^{th} vertex, and r_{mean} is the average of these vertex positions, or the centre of mass of these vertices, assuming each vertex is weighted equally. We would expect that as walks get longer, the radius of gyration of those walks will increase on average. For walks of N steps, we have the famous scaling relation from Flory [13]:

$$R_g \sim N^{\nu} \tag{1.2}$$

where ν is known as the *Flory exponent*. It turns out that for non-self-avoiding walks in 3-space like the ideal chain, $\nu = 0.5$, whereas in self-avoiding walks $\nu \approx 0.588$ [13]. In other words, we expect non-self-avoiding walks to be more compact in general than self-avoiding walks of the same length. Later we will be confining random walks and we will see that limiting the radius of gyration of a walk has important implications for its knotting.

1.2.1 Knots in random walks

The presence of knots in polymer systems has been predicted and shown to affect their physical properties [2, 3, 7, 8]. For example, long relaxation time modes appear in knotted polymer configurations which are not seen for unknotted configurations [60]. They also have interesting effects on the stretching of DNA molecules, making them more difficult to stretch, and applying a hysteresis like effect to the coiled-stretched transition [61, 62]. A remarkable mechanical property of knotted DNA was simulated where the DNA was ejected from a viral capsid, with a knot inside the capsid acting as a ratchet, ensuring efficient ejection [63]. Knots are also predicted to affect the opacity of polymers [64] with more complex knots being more opaque.

In the early 1960s it was conjectured by Frisch, Wasserman [9] and Delbrück [10] that all sufficiently long polymers are knotted. Early numerical results from Vologodskii et al. [65] using random walks as models for polymers supported this notion. While the lengths reached were much limited by computing power of the day, an increase in knotting probability was seen with walk length. In 1979, Kendall proved that every infinitely long Brownian walk contains infinitely many knots [66]. However, these same Brownian walks contain infinitely many self-intersections.

In the late 1980s it was finally proved independently by Sumners and Whittington [11] and Pippenger [12] that a self-avoiding random polygon on the cubic lattice has a probability of knotting approaching 1 as length tends to infinity. In particular, it was shown that this approach was exponential in form:

$$P(N) \sim 1 - e^{-N/N_0} \tag{1.3}$$

where $P(N)$ is the probability that a walk of N steps is knotted and N_0 is some characteristic length scale for knot proliferation, which will depend on the properties of the walk. Numerical evidence supporting this was soon to arrive [67, 68, 69], finding that the value of N_0 can vary over many orders of magnitude depending on the flexibility of the walk and the solvent quality. Proofs later followed that this exponential form holds true also for Gaussian random walks [70] and ideal walks [71] off-lattice.

It is natural to ask how the probability of finding a given knot varies with the length of the walk, or how the proportions of different types of knots vary. It is no surprise that the complexity of the knots seen increases with the length of walks and indeed that walks with multiple knots (such as two trefoil knots

or a trefoil and a figure-eight) dominate at large lengths [70]. Knots which are distinct from their mirror image, called *chiral* knots, are also seen to dominate as length increases [70]. The form of how the probability of a given knot varies has been investigated thoroughly [72, 73, 74, 75, 76] and it has been found to rise first polynomially, before decaying exponentially. The rate of this rise and fall varies with knot type, with simpler knots rising first, at shorter lengths, before giving way to more complex knots at longer lengths.

The presence of knots has a remarkable effect on the size of the walks. Earlier we noted that the Flory exponent for R_g scaling in ideal chains is $\nu = 0.5$. It was theoretically predicted that if only random walks of a given knot type are considered, they ought to scale as self-avoiding walks, with $\nu \approx 0.588$ [77]. This has received significant attention since, and appears to be confirmed numerically for both open and closed walks [59, 78, 79, 26, 80, 81] and has been termed topological swelling. The apparent paradox that the same random walk model can scale in two different ways is resolved by considering the knot types that become more likely as the walks become longer. While any given knot type will scale as $\nu \approx 0.588$, the average R_g of knots is smaller the more complex they are. As the walks get longer and more complex knots become more common, the decreased size of the complex knots balances the self-avoiding growth of any specific knot type [78]. Walks have to achieve a certain length before this trend is seen however, and there is indication that this length is significantly larger than the length at which knotting becomes more likely than unknotting.

It is seen in loose, unconfined polymers and random walks that knots tend to localise in small sections of the chain [82, 83]. An interesting effect in compact polymers and walks, either in confinement or in poor solvent conditions is to delocalise the knot throughout the curve [25, 84, 85, 86, 87, 88]. This is likely for entropic reasons, although there remains a question of why the localisation effect in loose configurations is so apparent while the free energy cost of a more delocalised knot is relatively low [88].

There is great interest in investigating polymers in spherical confinement with applications to spherical micelles [89] and most prominently DNA inside viral capsids [90, 91, 92, 93, 63, 94, 95]. Naturally, this confinement has consequences for the knotting of polymers and the random walks which model them. Many of the effects of tightening confinement mirror the effects seen for increasing length. Unsurprisingly, a tighter confinement leads to more knots, and more complex knots [96, 97] as well as more composite knots [98]. Furthermore,

if the probability of a given prime knot is traced as confinement is tightened, the probability first rises as unknots become less likely, and then decreases as more complex knots become more likely [98, 75].

Other confinement shapes have also been looked at, including tubes which act as models for microfluidic channels and membrane pores [99, 100, 101] and slits (two parallel walls), an environment similar to thin film conditions and other microfluidic cases [102, 103, 104, 105]. Changing the shape of confinement has important consequences for knotting. Many of the same results as for spheres hold to a point. However, as these geometries are tightened, the probability of knotting in general goes through a maximum before decaying once more [102, 105].

Most of what we have said here pertains to closed random walks. However, in this thesis we will be looking at open random walks, and we might wonder how many of these closed results apply to open walks. In walks where knots are localised the knotting statistics of open and closed walks are likely to be very similar [106], and much of what we have said here will be equally applicable. This has been studied directly, for example, by Millett, Dobay and Stasiak [26] using unconfined random walks. They took the dominant knot across closures to represent the knotting of the curve, finding that their knotting statistics were close to those of closed walks. They did identify a small number of walks where no single type of knot dominanted, but this number was too low in their unconfined system to worry about further.

We will be looking at confined open walks though, where we expect knotting to be delocalised and potentially different from closed walk knotting. Tubiana, Orlandini and Micheletti [28] studied open walks of varying degrees of compactness, comparing two different methods of knot recognition. They found that the two methods agreed which knot was tied for loose walks, but differed often for compact walks, writing that the transition to a compact state signals, 'a non-trivial increase of the geometrical complexity of confined polymer rings.' It is this complexity not found in closed curves that we will investigate and quantify using our virtual knot method.

1.3 Thesis overview

This thesis is an exploration of using virtual knots to analyse the knotting of open space curves, with proteins and random walks investigated specifically as

important systems with much historical and present interest. In Chapter 2 we will provide the necessary mathematical background to understand the thesis. This will cover the foundations of classical knot theory including knot types, ways of manipulating knot diagrams by way of Reidemeister moves, Gauss codes and distinguishing knots using knot invariants. We then discuss virtual knots in a more formal manner, their similarities and their differences from classical knots. This will include different interpretations of virtual knots, how virtual knots may be manipulated and virtual knot tables. Particular attention will be paid to virtual knots which can be used to interpret open curve projections, which we call minimally genus one virtual knots. We will go into more detail regarding different closure schemes, and briefly mention slipknots, which are another important geometrical aspect of open curve knotting.

With this established, we introduce the specifics of how we detect knots in open curves in Chapter 3. This is the foundation for all the numerical analysis undertaken in the rest of the thesis. We detail how we reproduce previous results which involve multiple physical closures of the space curve through a process we call sphere closure, as the curve is closed to many points uniformly distributed on a large sphere surrounding the curve. We then describe how we use virtual knots to recognise knotting through a process we call virtual closure. This takes projections of the curve from the same uniformly distributed points as in sphere closure and classifies these as virtual knots. We describe how we use knot invariants to distinguish both classical and virtual knots, providing worked examples for all invariants used. We investigate how the number of closures taken affects the spectrum of knots seen in both sphere and virtual closure, seeing that the statistics stabilise around the commonly used 100 closures.

We then undertake an initial exploration of the features one would expect to see under virtual closure. An important distinction we make here is between strong and weak knots. Strong knots are open curves which present the same type of knot in a majority of closures, whereas in weak knots, while the unknot is still less common than non-trivial knots, no single type of knot covers a majority of closures. We discuss how we expect weak and strong knotting to differ between sphere and virtual closure, anticipating that weak knotting will be more prevalent under virtual closure. We cover how knot type may vary as closure direction is changed, with the complete picture of knotting on closure in every direction being captured in what we call a knot globe. This leads us to a discussion about knot space, which we use here to refer to the different knot

types that may be obtained from single crossing flips in a knot diagram. Virtual knots given their in-between property lie in the spaces between classical knot types. We provide a diagram of the shape of knot space up to 7 crossings.

With the necessary machinery then in place, we look at knotting in proteins in Chapter 4. We begin with some more background about the experimental methods used to probe protein structure, the Protein Data Bank (PDB) is introduced as the online repository in which all protein structure information is kept and some of its limitations are covered. Then we detail how we perform our own survey of knotting in the PDB using virtual closure, including which chains we analyse and how we parse their atomic coordinates. We present the results of our investigation, finding a 25% increase in the number of knotted protein chains under virtual closure compared to sphere closure. Of these, we find around 40% are weakly knotted, a much higher fraction than under sphere closure. We cover some of the geometrical features of proteins in the PDB and how this compares to knotted chains, and pick out some families of proteins showing distinct knotting characteristics.

Then, to put these results into more context and to explore situations where we expect the knot type of open curves to be particularly ambiguous, in Chapter 5 we undertake extensive numerical investigations of confined, open random walks as models of polymers. We explore both on-lattice random walks and off-lattice walks and describe how we generate the walks in some detail. The on-lattice walks are confined to cubes, while the off-lattice walks are confined to spheres, tubes and between two parallel planes, referred to as slits, and outside of confinement. In order to confine the off-lattice walks to tubes and slits we must extend work already done to confine such walks to the sphere, and we detail how we do this and how successful we are at producing a uniform distribution of walk vertices.

Using these walks, we explore how the length of walk and the size of the confining volume affect their knotting statistics. First we compare on-lattice walks in cubes and off-lattice walks in spheres and unconfined, finding that the confined off-lattice walks knot more often than confined on-lattice walks and unconfined off-lattice walks. However, we find that the confined walks are much more likely to be weakly knotted than the unconfined walks, with the on and off-lattice models both giving very similar results, despite their overall knotting differences. By focussing solely on off-lattice walks, we investigate a much broader range of parameters, finding that longer walks and tighter

confinement both lead to increased knot probability and increased weak knot probability. We define the degree of confinement as a measure of how compact the confined walks are compared to their size outside of confinement and find that, while overall knot probability depends both on degree of confinement and length, weak knotting depends almost entirely on degree of confinement. We extend our analysis then to tubes and slits, finding that at equivalent length and radius of confinement, walks in tubes are less frequently knotted and less frequently weakly knotted than walks in spheres, with this even more so the case for walks in slits. With a suitable degree of confinement definition, we see that the weak knotting probability of all our walks against degree of confinement lie on the same curve.

We conclude in Chapter 6 with a summary and consolidation of our findings, folding the protein and random walk results together to find that while proteins very rarely knot compared to random walks, they are comparably weak in their knotting. We reflect on virtual closure and its place among the many methods of knot recognition in open curves, including a discussion on the recently developed method of knotoid analysis and how this compares to the closely related virtual closure. In closing we discuss where this work could be taken in the future.

Finally, we include an appendix containing visualisations of the knot globe for specific random walks and proteins as examplars for the different types of knotting seen.

2 Background

In this chapter we will cover the mathematical background material necessary for future chapters. This will include some introductory knot theory, both classical and virtual and an overview of techniques for recognising knots in open curves. The intention is to introduce only that material which will be useful in every following chapter, leaving more specific detail to be explained where necessary.

2.1 Knot theory

Until recently, mathematical knot theory dealt only with closed curves. The challenge when analysing open curves is to adapt the powerful mathematical tools developed for closed curves to topologically trivial, but geometrically complex space curves. We will start with a review of classical knot theory, before moving onto the more modern innovation of virtual knots which will be more useful when trying to capture the complexity of open curves. Much of the content on classical knot theory is available in Adams [24], and further reading is available there.

2.1.1 Classical knot theory

Classical knot theory is the branch of knot theory which deals with the topology of closed curves. Strictly, a knot is an embedding of the circle, $\mathbb{S}^1$, into 3-space with no self-intersections. For most introductory applications it is sufficient to think of this space as regular $\mathbb{R}^3$, although for some more formal situations it is helpful to use instead the surface of a four-dimensional sphere, $\mathbb{S}^3$. When studying the topology of knots, we allow the embeddings to be deformed continuously as long as no cutting, glueing or self-intersections occur. Deformations

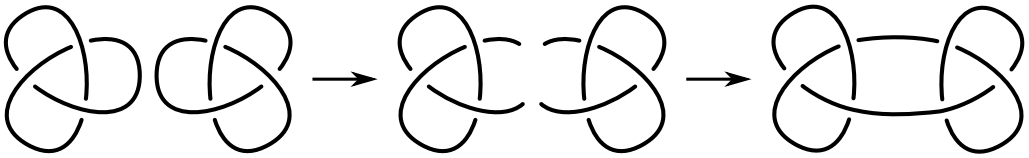


Figure 2.1: The process of taking a connected sum of two trefoils to produce a composite knot.

of this kind are called *ambient isotopies*. If two different embeddings can be transformed into each other by ambient isotopy, they have the same *knot type*. All embeddings of a particular knot type can be deformed into any other embedding of the same knot type, and none can change knot type under ambient isotopy. If an embedding is ambient isotopic to a planar circle, it is the trivial knot known as the *unknot*.

The *connected sum*, or *knot sum*, of two knots is created from two non-trivial knots far apart, such that there is no overlap or linking between them. A small section of each knot is removed, and the two now open curves are then joined together to create one closed curve. This process is illustrated in Fig 2.1. Knots created in this way are called *composite knots* and include the well known granny and reef knots which are composites of two trefoil knots. Knots which cannot be created from connected sums of other knots are called *prime knots*, and all composite knots have a decomposition into their component prime knots. Crucially there are no inverse knots i.e. there are no knots which, under a connected sum produce the unknot.

It is often useful to consider a projection of the knot to the plane $\mathbb{R}^2$, or the surface of a three-dimensional sphere $\mathbb{S}^2$. In order to preserve topological information, intersections of the projection with itself are decorated with *crossings* indicating which strand passes over which. Care must be taken in choosing a projection where no more than two strands intersect at a given point. The resulting decorated 4-valent graph is known as a *knot diagram*. All the topological information of a closed space curve is contained in its knot diagram, regardless of how the curve was projected. An *oriented* knot diagram is one in which a direction of circulation is chosen on the knot diagram. This is shown in Fig 2.2, where a space curve is projected and an orientation applied to the resulting knot diagram. The relative directions of circulation at a crossing give each crossing a *sign*, either positive or negative as shown in Fig 2.3, which does not depend on the orientation chosen. The *writhe* of a knot diagram is the sum

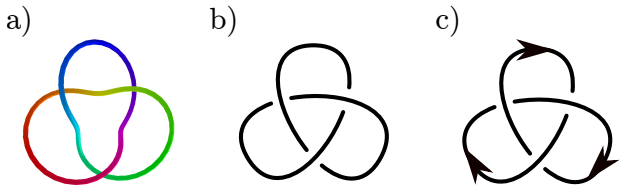


Figure 2.2: a) a knotted closed curve in 3-space. b) The knot diagram of this curve on projection. c) An oriented knot diagram.

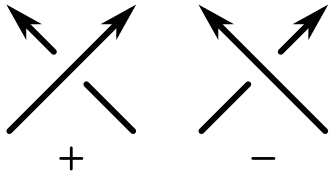


Figure 2.3: The sign of a crossing in an oriented knot diagram.

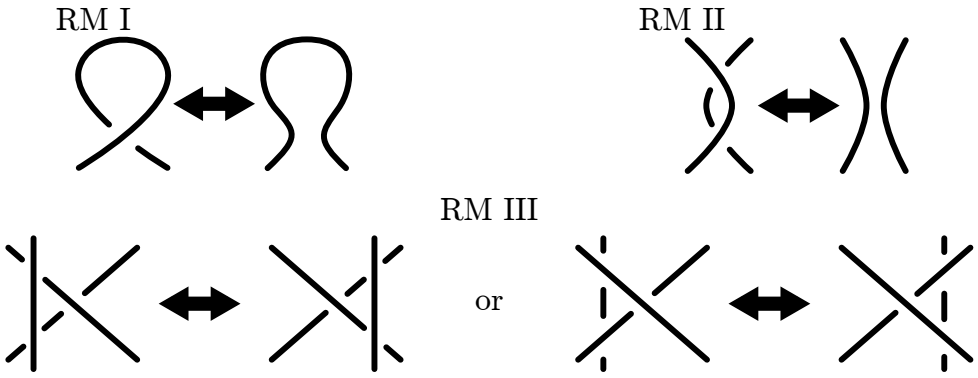


Figure 2.4: The three classical Reidemeister moves which are used to manipulate knot diagrams.

of the crossing signs.

Knot diagrams can be manipulated in analogous ways to space curves. *Planar isotopy* is the continuous deformation of strands of the diagram without cutting or glueing. Crossings may not be altered by planar isotopy however, so we require additional moves in order to capture the possible space curve manipulations. Only three such moves, known as the *Reidemeister moves*, shown in Fig 2.4, are needed to manipulate any knot diagram to any other knot diagram of the same knot type. The first Reidemeister move untwists a loop, removing a crossing. The second moves two overlapping loops away from each other, removing two crossings. Each of these may be reversed to add crossings. The third involves moving a strand past a crossing, leaving all crossings intact.

The *minimum crossing number* of a knot is the least number of crossings

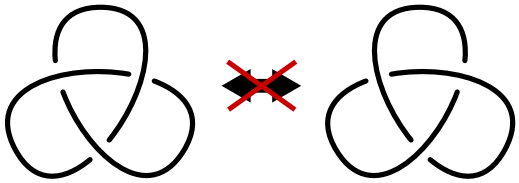


Figure 2.5: The trefoil knot is chiral. This means that it is not possible to deform a trefoil knot into its mirror image.

minimum crossing number	0	3	4	5	6	7	8	9	10	11
number of prime knots	1	1	1	2	3	7	21	49	165	552

minimum crossing number	12	13	14	15	16
number of prime knots	2,176	9,988	46,972	253,293	1,388,705

Table 2.1: Number of prime knots of a given crossing number. Chiral pairs are counted as one knot type.

needed to represent it in a knot diagram. Such a *minimal diagram* may always be reached with Reidemeister moves and planar isotopy, although crossings may need to be added during the transformation. Knot types are labelled n_m where n is the minimum crossing number and m an arbitrary label used to distinguish distinct knots with the same minimum crossing number. Some knots cannot be transformed into their mirror image, such as the trefoil shown in Fig 2.5. These knots are called *chiral* and both receive the same label. The number of knots of a given minimum crossing number grows rapidly, although how rapidly is still an open question. The number of prime knots of a given crossing number is given in Table 2.1, where chiral pairs are counted as one knot. The prime knots are often gathered together in *knot tables*, arranged according to their labels, as in Fig 2.6. These tables are only complete up to knots of 16 crossings [107].

All the topological information about a knot is contained in the crossings of its knot diagram. There are a number of ways of encoding this crossing information, such as Dowker notation and Conway notation. *Gauss code* is a particularly clear method and is used most extensively in this thesis. To produce the Gauss code, first a base point is chosen somewhere along the knot diagram.

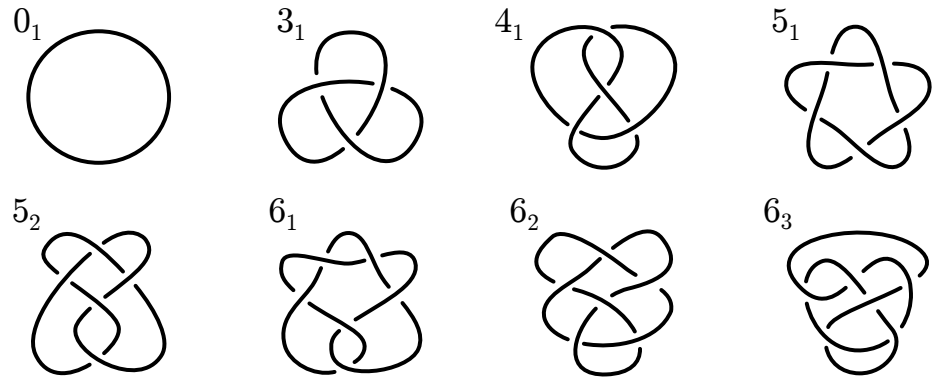


Figure 2.6: A knot table of the classical knots up to six crossings, with their usual labels.

Then the diagram is traversed in one direction and the crossings labelled from 1 in the order they are encountered. When encountering a crossing, the crossing label is recorded along with whether the current strand passes over or under, and the sign of the crossing. Once a complete circuit of the knot diagram has been made, the Gauss code is complete. As this contains the complete information about the relative positions of crossings, the knot diagram may be recovered from the Gauss code, up to planar isotopy. The Gauss code for the trefoil shown in Fig 2.6 is +1o, +2u, +3o, +1u, +2o, +3u, where we have taken the top most point as the base point and proceeded clockwise. The sign of the crossing is given first, followed by the crossing label and then whether the current strand passes over or under, denoted 'o' and 'u' respectively.

While any knot may be represented as a knot diagram, and any knot diagram may be reduced to a minimal diagram from which the knot type can be identified, it is not practical to go through this process every time the knot type of an arbitrary space curve is to be found. *Knot invariants* are quantities that can be calculated from a knot diagram which do knot vary under planar isotopy or Reidemeister moves as they depend only on knot type. If two knots have different knot invariants, they are different knot types. However, depending on the particular knot invariant compared, distinct knots may have the same invariant. For example, the minimum crossing number is a knot invariant, but it cannot distinguish 5_1 from 5_2. Fortunately, more powerful knot invariants have been found, many of which are easier to calculate in general than the minimum crossing number.

Many practically useful knot invariants take the form of polynomials. An early example is the *Alexander polynomial*, $\Delta(t)$, which is simple to calculate and

performs well for simple knots, distinguishing all knots up to 8 crossings [108]. It cannot however distinguish the handedness of chiral knots. For composites knots, the resulting Alexander polynomial is the product of the Alexander polynomials of the prime knot factors.

The calculation of the Alexander polynomial for a given knot diagram increases as the square of the number of crossings. This calculation may be sped up, at the cost of some discriminating power, by calculating $\Delta(-1)$ as opposed to the full symbolic polynomial. The resulting integer invariant is known as the *determinant* and is a practical alternative for identifying simple knots from complex diagrams.

Later, the *Jones polynomial*, $V(q)$ was found, which is a more powerful knot invariant, distinguishing more knots from each other [109]. Of note is that the Jones polynomial can distinguish some chiral knots. It is still an open question whether it can distinguish the unknot from all non-trivial knots, unlike the Alexander which demonstrably cannot. Unfortunately, this comes at the cost of an exponential increase in computing time with crossing number. Calculating $V(1)$ is a useful efficiency saving here also, with a similar trade-off in power as the determinant, and composite knots similarly have Jones polynomials which are products of their prime factors.

Shortly after the discovery of the Jones, the *HOMFLYPT* polynomial was found using a similar technique [110, 111]. Sometimes called the HOMFLY polynomial or the generalised Jones polynomial, it is a polynomial of two variables, denoted $P(l,m)$, and takes longer to calculate than the Jones but still scales exponentially with crossing number. The HOMFLYPT is strictly more powerful than the Alexander and Jones as both polynomials can be recovered from the HOMFLYPT. $\Delta(t) = P(1, t^{1/2} - t^{-1/2})$ and $V(q) = P(q^{-1}, q^{1/2} - q^{-1/2})$. Accordingly, the HOMFLYPT polynomial of composite knots shares the same relation to their prime factors as the Jones and the Alexander.

So far, all discussion has regarded the analysis of a single closed space curve but many of these concepts are directly applicable to the generalised case of multiple curves, called *links*. As links will not feature in this thesis, more detail shall not be presented.

2.1.2 Virtual knot theory

Virtual knot theory, introduced by Louis Kauffman in 1996 [29], extends classical knot theory by allowing a new type of crossing called the *virtual crossing*, which

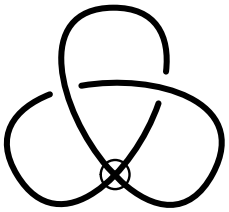


Figure 2.7: A virtual knot, in particular the virtual trefoil or $v2_1$. Note the virtual crossing indicated by the circle.

is marked with a circle as in Fig 2.7. Technically, all classical knots are also virtual knots but for our purposes here we will only call knots whose minimal diagrams contain at least one virtual crossing *virtual knots*. Many of the concepts from classical knot theory have analogues in virtual knot theory, some of which are directly applicable.

All the classical Reidemeister moves survive but there are also additional *virtual Reidemeister moves* which allow virtual crossings to be manipulated, as shown in Fig 2.8. Virtual Reidemeister moves one and two are essentially the same as their classical counterparts with classical crossings replaced with virtual ones. The third classical Reidemeister move has two virtual variants; the first involves passing a virtual loop over a classical crossing, and the second involves passing over virtual crossing. Additionally, there is a third variant of the third Reidemeister move called the *forbidden move* which passes a classical loop over a virtual crossing. If the forbidden move were allowed it would be possible to change the knot type of a virtual knot, possibly making it trivial where before it was non-trivial. It is also worth noting here that the crossing number of a virtual knot diagram is given by the number of classical crossings.

While the virtual knot diagrams were the original presentation of virtual knots, there are other interpretations of virtual knots. One may imagine a virtual knot as a space curve embedded in a thickened torus, $\mathbb{T} \times \mathbb{I}$ as in Fig 2.9. In less mathematical language, a thickened torus can be thought of as the space between two tori, each sharing the same centre hole and one fully enveloping the other, creating a toroidal shell. In this picture, a virtual knot diagram is created by projecting the space curve with classical crossings occurring when strands pass over each other on the same side of the torus, and virtual crossings occurring when strands pass on opposite sides of the torus. In the same way that the classical Reidemeister moves capture the aspects of ambient isotopy not covered by planar isotopy, the virtual Reidemeister moves capture ambient

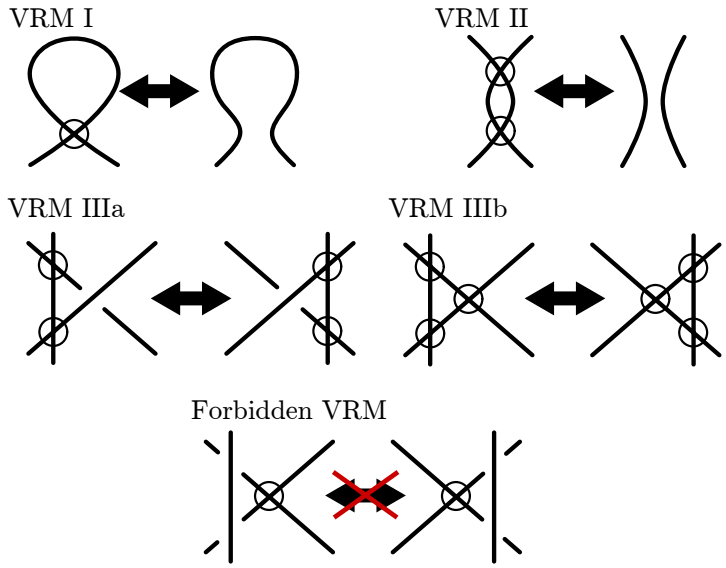


Figure 2.8: The virtual Reidemeister moves which are used to manipulate virtual knot diagrams, in addition to the classical Reidemeister moves. Also included is the forbidden virtual Reidemeister move with which virtual knot type may be changed.

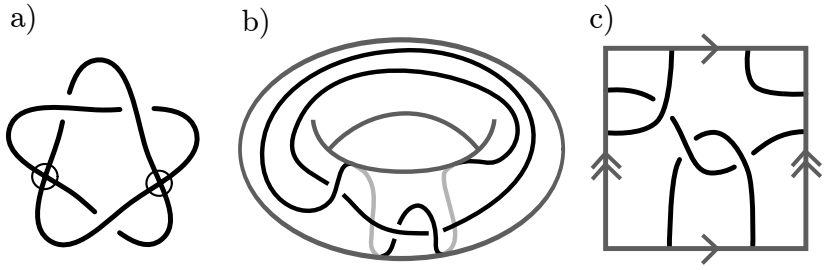


Figure 2.9: The virtual knot $v3_7$ as a) a knot diagram, b) embedded in a thickened torus and c) a different presentation of the embedding in a thickened torus. Note that virtual crossings correspond to where strands on opposite sides of the torus cross in projection.

isotopic manipulations in the thickened torus.

In order to generate some virtual knot diagrams, it is necessary to embed the space curve in a thickened surface of higher genus than a torus. The *genus* of a virtual knot is the lowest genus thickened surface a space curve can be embedded in to give the virtual knot on projection. All classical knots have a genus of zero as they can be embedded in a thickened sphere, but if a diagram has any virtual crossings it must have at least genus one.

The Gauss code of a virtual knot diagram is found in much the same way as a classical knot diagram, by choosing a start-point and orientation and recording

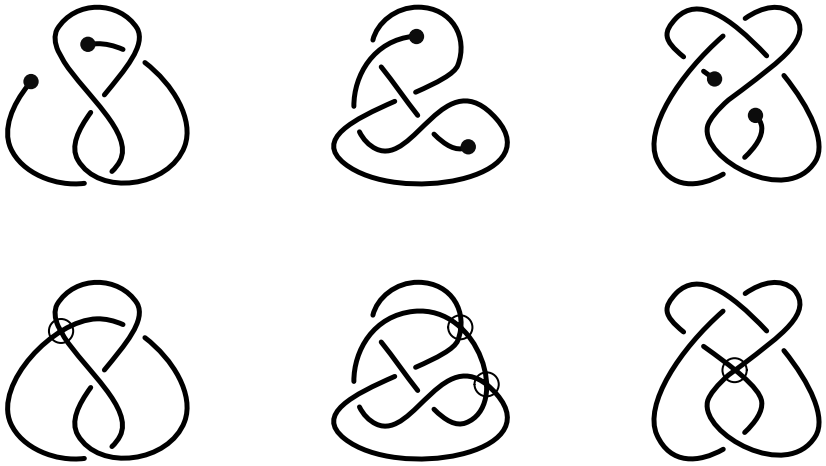


Figure 2.10: Projections of open curves (top) and the minimally genus one virtual knots which share the same Gauss code (bottom). The end-points of the projected open curves are picked out with dots.

crossing information encountered in a circuit of the diagram. However, virtual crossings have no sign, and are not numbered as the classical crossings are. The Gauss code does not contain explicit information about the virtual crossings and their position is implied. Kauffman [29] justifies this saying, '... the idea is not that a virtual crossing is just an ordinary graphical vertex. Rather, the idea is that the virtual crossing is not really there.' As an example, the Gauss code of the virtual knot in Fig 2.7 is −1u, −2o, −1o, −2u, if we start in the lower left lobe and proceed clockwise.

This suggests a further interpretation of virtual knots as projections of open curves. If one generates the Gauss code of a projection of an open curve with one of the ends as the start-point and stopping at the other end, the Gauss code may not correspond to a valid classical knot diagram, but it will correspond to a valid virtual knot diagram. Examples of projections of open curves and the virtual knots which share the same Gauss code are given in Fig 2.10. However, only certain genus one virtual knots may be generated in this way. We define *minimally genus one* virtual knots to be those virtual knots whose diagrams may always be deformed such that all virtual crossings lie on a single arc, starting and stopping at classical crossings. Only minimally genus one virtual knots can result from projections of open curves. The knot in Fig 2.7 is a minimally genus one virtual knot, while the knot in Fig 2.9 is genus one, but not minimally genus one.

Chirality exists in virtual knots also, although there are more types of chirality

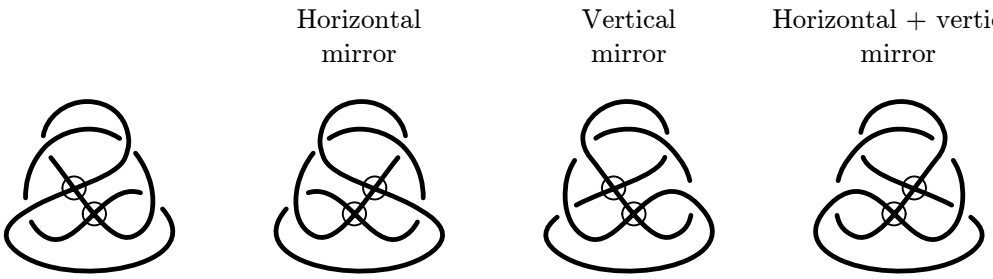


Figure 2.11: The virtual knot $v4_{36}$ and its horizontal, vertical and combined horizontal and vertical mirrors.

than with classical knots. The *horizontal mirror* is the equivalent of chirality in classical knots, where there is a reflection of the virtual knot diagram in the plane, preserving over and under crossings. This has the effect in the Gauss code of inverting crossing signs, + to − and vice versa, and just as with classical knots, virtual knots may or may not be equivalent to their horizontal mirrors. *Vertical mirrors* instead flip over and under crossings. With classical knots, this is equivalent to taking a projection of a closed space curve from the opposite direction; as the space curve topology has not changed, the topology of the knot diagram has not either. However, the same is not true of virtual knot diagrams and there exist virtual knots whose vertical mirrors are not equivalent. Fig 2.11 shows a virtual knot which is not equivalent to its horizontal or vertical mirrors.

An additional consideration which does not affect classical knots, although it can affect classical links, is that of inverses. An *inverse* of a knot diagram is one in which the orientation of the diagram is reversed; that is, the direction of circulation about the knot used to endow the crossings with a sign. While all classical knots may be deformed to their inverses, this is not true of all virtual knots.

Knot tables have been constructed for virtual knots, although they are not as developed as their classical counterparts. Jeremy Green, under the supervision of Dror Bar-Natan, generated a virtual knot table up to 6 classical crossings, although knot diagrams have only been drawn for knots of up to 4 crossings [112]. It is from this table that we take the labels we use for virtual knot types. In the original table these are given in the same format as for classical knots, n_m, where n is the number of classical crossings in the minimal diagram and m is an arbitrary label. In order to distinguish virtual knots from the familiar classical knot labels we prefix the label with a v, as in vn_m. Fig 2.12 is a knot table of all the known minimally genus one knots up to four crossings with the labels used

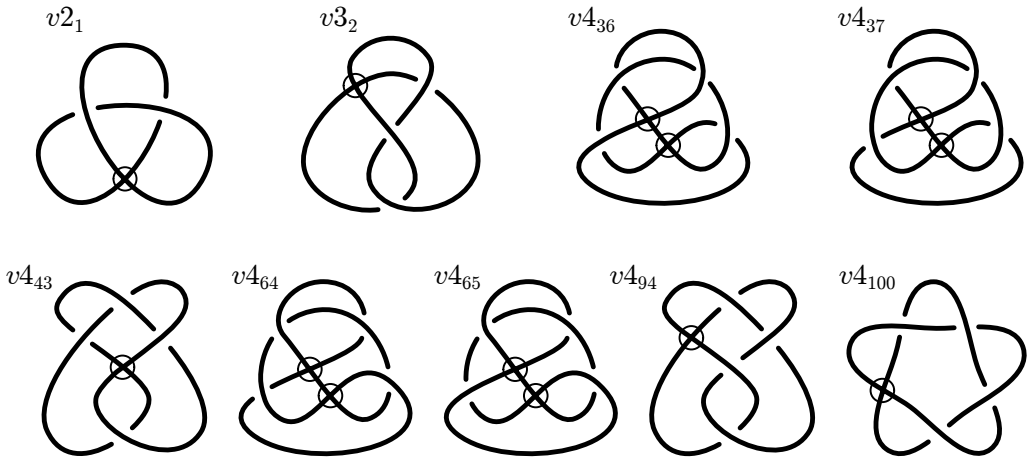


Figure 2.12: The prime minimally genus one virtual knots up to four classical crossings.

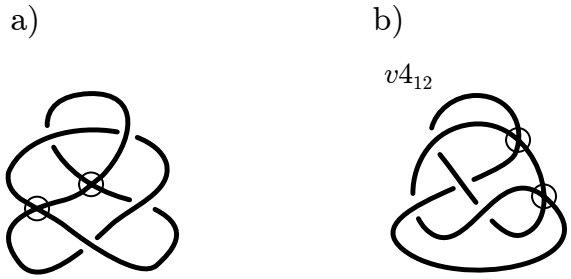


Figure 2.13: In our paper [113] we used the diagram a), to represent $v4_{94}$. This is not wrong, but the presentation used in Fig 2.12 is simpler. We also claimed the knot $v4_{12}$ was a prime virtual knot, but it was pointed out to us by the authors of [114] in private communication that this is in fact a composite of two $v2_1$ knots of opposite handedness.

in Green's table. See also Fig 2.13, which indicates the corrections made to the virtual knot table included in our publication [113] to arrive at the table shown in Fig 2.12. The number of virtual knots of a given crossing number in Green's table are given in Table 2.2. It is clear to see from here that there are many more virtual knots of a given crossing number compared to classical knots. Virtual composite knots can be constructed in the same way as classical composites, and may be composites of a classical and a virtual knot, or two virtual knots.

In order to distinguish between these many virtual knots, *virtual knot invariants* have been devised, many with classical analogues. A quick to calculate, although not particularly powerful invariant is the *self-linking* number [115]. This is the sum of the signs of all the *oddly intersticed* crossings. A crossing is oddly intersticed if an odd number of other crossings are encountered between

minimum crossing number	2	3	4	5	6
number of oriented virtual knots	2	22	590	18,202	707,025
number of unoriented virtual knots	2	14	325	9,226	354,673
number of reduced virtual knots	1	6	107	2,442	90,232

Table 2.2: Number of virtual knots of a given crossing number in Green's table. Unoriented virtual knots treat inverses as equal, and reduced virtual knots treat inverses and horizontal and vertical mirrors as equal. Some composite knots are included in Green's table, but it is not clear if this is exhaustive. These numbers should be taken as a rough guide.

its over and under crossings. The self-linking number of all classical knots is zero as no crossings in a classical knot are oddly intersticed, which makes it a useful first check to determine if a knot is virtual. This is not foolproof however, as some virtual knots have a self-linking number of zero also, so we require stronger invariants.

There are a number of Alexander type invariants for virtual knots, but the only one we will use is the *generalised Alexander polynomial*, $\Delta_g(s,t)$ [116, 117]. This is a two variable extension of the Alexander polynomial which is zero for all classical knots. While it can be zero for virtual knots, this happens only rarely for the simple virtual knots which will be most common in the results of this thesis. The calculation time scales as the square of the crossing number, as the classical Alexander polynomial does, although as the matrix whose determinant must be calculated is $2n \times 2n$ as opposed to $(n-1) \times (n-1)$, it is slower at equal crossing number. While the generalised Alexander polynomial distinguishes many more knots than the self-linking number, it fails to distinguish the two simplest minimally genus one knots, $v2_1$ and $v3_2$.

Perhaps surprisingly, the Jones polynomial can be calculated for virtual knots in exactly the same manner as for classical knots and remains an invariant. As the calculation does not change, it scales exponentially exactly as for classical knots. Crucially for practical purposes, the Jones polynomial does distinguish the two simplest minimally genus one virtual knots.

In constructing the table of minimally genus one virtual knots, in addition to the table of Green and Bar-Natan [112] we drew on the work of Andreevna and Matveev, who tabulated the genus one virtual knots up to 5 crossings [114]. We determined the minimally genus one cases by inspection and found there were two additional knots in the Andreevna table, compared to the Green table. By calculating the generalised Alexander and Jones polynomials, we managed

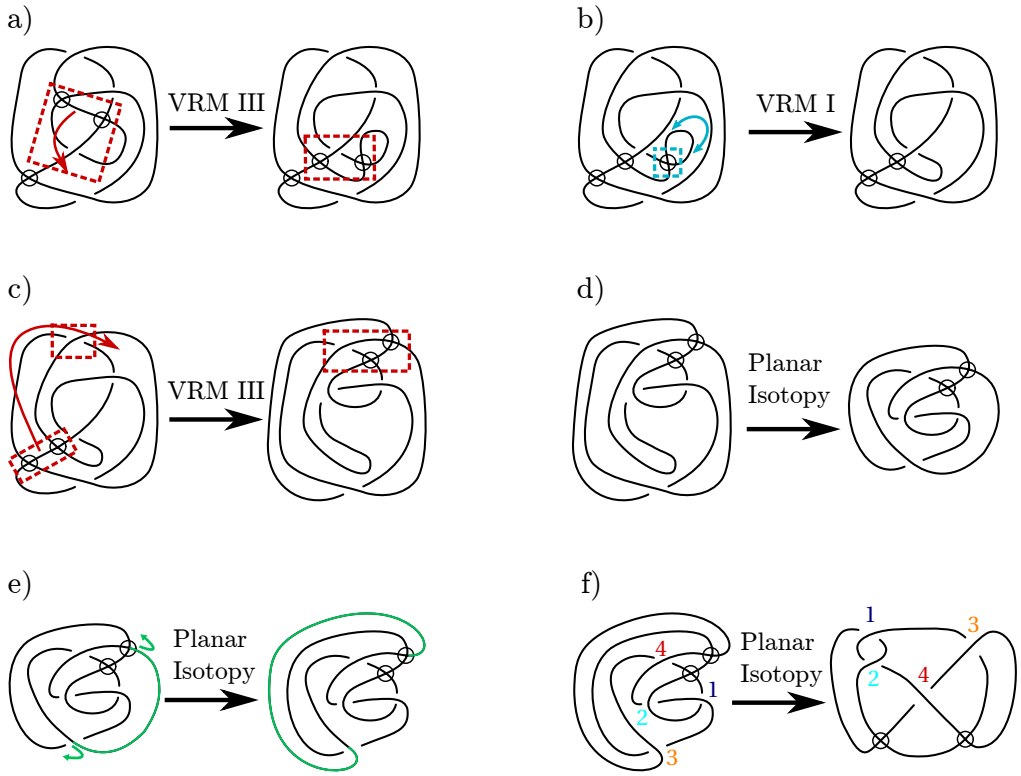


Figure 2.14: Transformation between two depictions of the virtual knot $v4_{64}$. The initial conformation in a) is that depicted as 4_{64} in the virtual knot table of Green and Bar-Natan [112]. An alternative presentation of this knot from the genus one table of Andreevna and Matveev [114] (labelled there as 4_8), is shown in f). b)-e) show how a) may be transformed to f) by a combination of virtual Reidemeister moves and planar isotopies of the knot. In e), the planar isotopy moving the green strand across the knot is not directly allowed by the virtual Reidemeister moves, but as the knot diagram is implicitly drawn on $\mathbb{S}^2$ this represents the strand passing 'behind' the sphere (or on the plane, passing through infinity). Analogous moves can be made to transform Green's 4_{36} to the vertical mirror of Andreevna's 4_9 which share the same skeleton but with different under and over-crossings.

to determine that the knot labelled 4_{36} in Green's table has equal invariants to the vertical mirror of the knot 4_9 in the table of Andreevna, and Green's 4_{64} had invariants equal to Andreevna's 4_8. While Green's 4_{36} and 4_{64} do not appear to be minimally genus one by inspection, we found a sequence of Reidemeister moves which transform them into Andreevna's 4_9 and 4_8 respectively. These are given in Fig 2.14.

2.2 Knot detection in open curves

Open space curves, which can be thought of as the interval, $\mathbb{I}$, embedded in $\mathbb{S}^3$ or $\mathbb{R}^3$, are topologically trivial in that they can always be deformed to a straight line. Nevertheless, some open curve conformations bear a striking geometrical resemblance to knotted closed curves. As such, methods to measure this conformational commonality have been developed and will be discussed in this section. Also included is a note on slipknots, which are a purely geometrical feature of space curves which can have important physical consequences.

2.2.1 Methods of detection

Every method of knot recognition in open curves until recently has involved a closure of the curve in its original space. By joining the ends together, a closed curve is formed and can be analysed using the tools of classical knot theory. The simplest and most naive way of doing this is called *direct closure*, which connects the ends with a straight line. While the result of this closure may agree with the knot type we would intuit for some curves, it is not hard to think of examples where this closure path goes through a tangled section of the curve to produce a very different knot type [26, 27].

A more sophisticated closure method might seek to avoid these situations and perform the closure as far away from the bulk of the curve as possible. *Radial closure* involves taking a line from each of the end-points and heading directly away from the centre of mass of the curve, extending far enough away that the ends of these lines can be joined without the possibility of interfering with the rest of the original curve [118]. While this avoids some of the problems of direct closure, it carries with it its own flaws. For example, a curve whose ends are relatively close together, but are embedded deep within the rest of the curve may have closure lines extending in opposite directions through the curve, introducing complexity where a simple direct closure would have been more appropriate.

The *minimally-interfering closure* method marries these two schemes together by taking advantage of the *convex hull* of the curve [28]. The convex hull is the convex set of the vertices of the curve. For points in the plane, the convex hull takes the same shape as a rubber band stretched to enclose all the points. This can be extended to three dimensions to produce a surface surrounding the curve. The minimally-interfering closure method then, involves finding the

distance between the end-points, and comparing this to the shortest distances from each end-point to the convex hull. If the ends are closer to each other than to the convex hull, direct closure is used, otherwise lines are added to each end taking the shortest path to the convex hull and closed outside.

Another scheme developed by Taylor [21], sometimes referred to as a *primitive path* method [119, 118], involves smoothing the curve. The ends are kept fixed and the rest of the curve is 'contracted... as if it were a rubber band.' Topologically trivial curves will be reduced to straight lines, whereas knotted curves will be left with a tight, well localised knot far from the ends. The ends can then be joined together in an obvious manner without interfering with the knotted region.

Each of the above methods are often successful at recognising the same knots as one might do intuitively, particularly for simple open curve conformations. However, they only perform a single closure, returning a single knot type and provide no information on how confident they are in this identification. If a given open curve is particularly complex, these methods will still return a knot type as assuredly as they would a simple curve, even if a slight perturbation to the curve would change the recognised knot type drastically.

To capture the more ephemeral qualities of open curves, methods involving multiple closure have been proposed [20, 26, 27, 118]. Each of these involves closing the curve to points on a sphere of effectively infinite radius surrounding the curve as was shown in Fig 1.4. By taking many closures, the resulting ensemble of knots can be analysed statistically to find the closed curve knot which the open curve most resembles, as well as providing a measure of how close this resemblance is. The methods differ only in how they join the ends to the sphere: *single stochastic closure* chooses random points on the sphere and joins both ends to each point with straight lines; *double stochastic closure* chooses pairs of random points on the sphere, joining each end to a different point and closing with an arc on the sphere; and the *uniform closure method* chooses a set of points which cover the sphere uniformly, and joins the ends to each in turn. In the rest of this thesis we will refer to the uniform closure method as *sphere closure*.

As the number of closure points increases and the relative proportions of knot types in the ensemble stabilise, the methods reassuringly converge to the same spectrum of knots [118]. While the most common knot type is sufficient for many purposes, it is useful to have the full spectrum to analyse also, showing

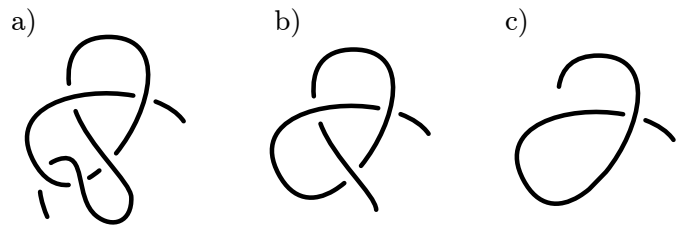


Figure 2.15: a) a projection of a slipknotted curve. By pulling the ends away from each other, an unknot is produced. b) after snipping away from one end, a trefoil knot is found. c) further snipping removes the trefoil, giving the unknot again. This is the characteristic under snipping of slipknots.

the various knots that the curve could become with some perturbation, with the relative fractions of each knot giving an idea of the size of perturbation that would be needed. With the increased computer power available today, the burden of calculating many knot invariants is not so great and so these methods have become more popular than single closure methods.

2.2.2 Slipknots

Slipknots are a relatively recent subject of study which have gained more attention as interest in knotted open curves has increased. They are a purely geometrical feature of open curves with interesting mechanical properties. A familiar example from every day life is of a shoelace bow, which is secure and holds together until the ends are pulled apart, resulting in the knot unravelling. The closure of such a string would result in an unknot and so would be topologically trivial even by the methods outlined above.

King, Yeates and Yeates propose a method to capture this feature of open curves [45]. To start with, the knot type of the open curve is determined using one of the above methods. Then, the curve is progressively shortened from one end, and the knot type analysed at each step. The essential feature of a slipknot, is that at some point the knot type will transition from one knot type, to a more complex knot type, and then back. For simple curves this is often a transition from unknot, to trefoil and then to unknot again, as illustrated in Fig 2.15. To fully capture the slipknotting of a curve, the curve must be shortened from both ends and to all possible degrees. The resulting information can then be presented in a triangular matrix showing the location of all knots and slipknots within a given curve. This technique has been used most notably by KnotProt, the knotted proteins database [23].

3 Detecting knotting in open curves using virtual knots

The previous chapter dealt in part with the difficulties of detecting knots in open curves, as well as a number of methods which have been used to overcome these difficulties. Here we detail a new method of knot recognition in open curves which follows from an original suggestion by Dr. Alexander Taylor. The key insight here is that the Gauss code of a projected open curve always corresponds to a virtual knot, which makes virtual knots a natural object to use when determining the knottedness of an open curve. This chapter describes how we use virtual knots to detect knotting in a process we call virtual closure. We start at a conceptual level and then go into the practical details of the complete knotting analysis used in the later parts of this thesis including how we calculate knot invariants. We then explore the virtual closure method in some detail, covering how we can categorise knotting according to the composition of the virtual knot spectrum of an open curve, the detail of 'knot globes' and implications for slipknotting. Also included is a picture of knot space, showing which classical knots are related to which by reversing crossings and which virtual knots lie between them.

3.1 Virtual closure

Of the methods of knot recognition detailed in Chapter 2, those which involve taking multiple closures of the open curve capture the most detail of the curve. All of these require adding sections to the space curve and analysing the resulting closed curves, and so the original curve is not analysed directly. To try to access information about the curve's conformation more directly, we propose a new method. Instead of physically joining the ends of the curve, we take projections of the curve and connect the ends in the projection, adding virtual

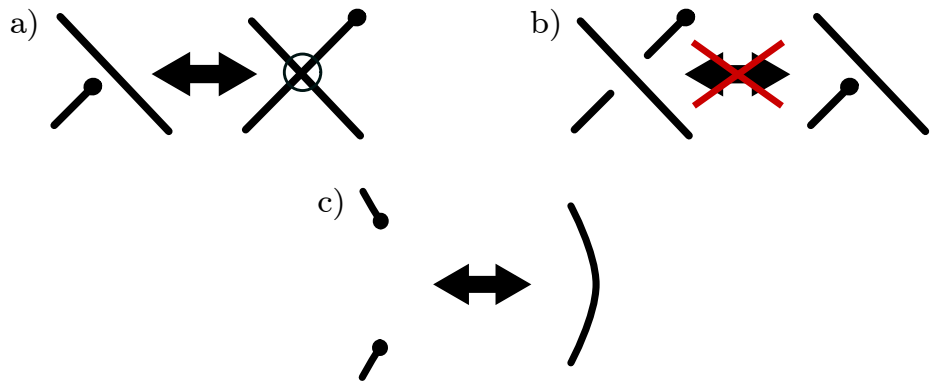


Figure 3.1: a) A valid manipulation of an open knot diagram during virtual closure which adds or removes a virtual crossing. b) An illegal move which may change the knot type of a subsequently closed diagram by adding or removing classical crossings. c) The final stage of a closure where the ends are joined.

crossings where strand crossings occur. We can then analyse these closed projections as virtual knots. Just as in sphere closure, where the curve is joined to points uniformly distributed on a large surrounding sphere, we project from uniform directions around the curve and obtain an ensemble of virtual knot types. This process we call *virtual closure*. Remember that the Gauss code of a virtual knot does not include information on the virtual crossings. In practice, when we analyse the Gauss code of a virtually closed curve, we are just analysing the Gauss code of the projected curve with no added information.

Not every virtual knot can occur on virtual closure, only those which we call minimally genus one virtual knots (see Sec 2.1.2). These virtual knots can be deformed such that their virtual crossings all lie on one arc. There is also no reason why a classical knot cannot be produced from virtual closure, if there are no strands in between each end point on projection, or if they cannot be removed by (virtual) Reidemeister moves. When manipulating projections for virtual closure, in addition to the classical and virtual Reidemeister moves, it can be helpful to make explicit the move depicted in Fig 3.1 a), where an end may be passed over an existing strand so long as a virtual crossing is added at the intersection. It should also be made clear that the ends may not be passed back through a classical crossing, removing it as in Fig 3.1 b), as this could potentially change the topology of the projection once virtually closed, the final stage of which is shown in Fig 3.1 c).

Virtual closure provides us with a new way to visualise the virtual Reidemeister moves. Each of the moves can be thought of as taking a different closure path

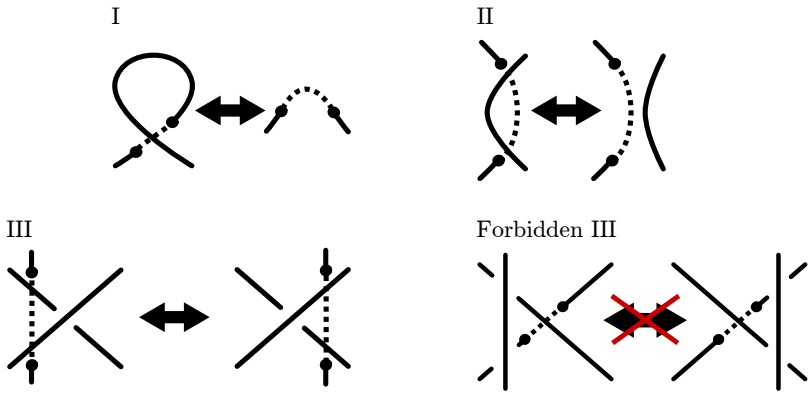


Figure 3.2: The three virtual Reidemeister moves have equivalents in the closure path taken during virtual closure. The dashed lines represent closure paths, and by taking different paths, the virtual Reidemeister moves may be recovered. The forbidden move is also included.

between the end points in projection. This may be made clearer by allowing the end points to move, but without performing a) or b) from Fig 3.1, and is shown in Fig 3.2. The relative abstraction of virtual crossings is avoided in these diagrams and the rationale behind the virtual Reidemeister moves can be seen. Additionally, the forbidden Reidemeister move is clearly illegal here is it violates the rule of Fig 3.1 b) by moving a classical crossing past an end point.

Putting this together, Fig 3.3 shows a simple open curve and three orthogonal projections. The panels to the right show the resulting knots upon virtually closing these projected diagrams. Two of the panels, the red and green, result in classical trefoil knots, whereas the blue panel gives a virtual trefoil knot, $v2_1$. While this example is relatively easy to evaluate by eye, a more complex curve like that in Fig 3.4 shows that this will not always be the case.

3.2 Methodological details

When fully analysing the knotting of an open curve in our own analysis, we use both virtual closure and sphere closure. These methods have been described conceptually elsewhere and we will cover the specifics of how we accomplish them here. The code we use to do the analysis takes as an input a series of (x, y, z) coordinates which are assumed to be connected in order, forming a piece-wise linear open curve.

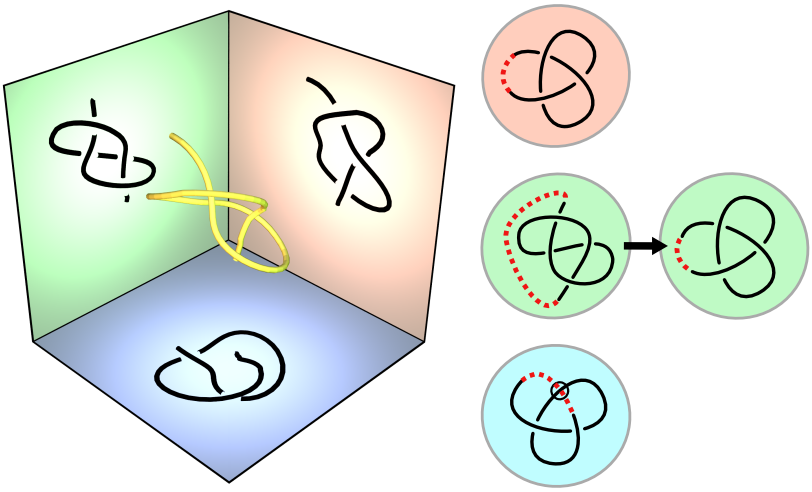


Figure 3.3: Orthogonal projections of a simple curve. The knot obtained on virtual closure is shown to the right, with the virtual closure taken along the dashed red line. In the case of the green panel, a few Reidemeister moves can be performed to show the trefoil knotting of this projection on virtual closure.

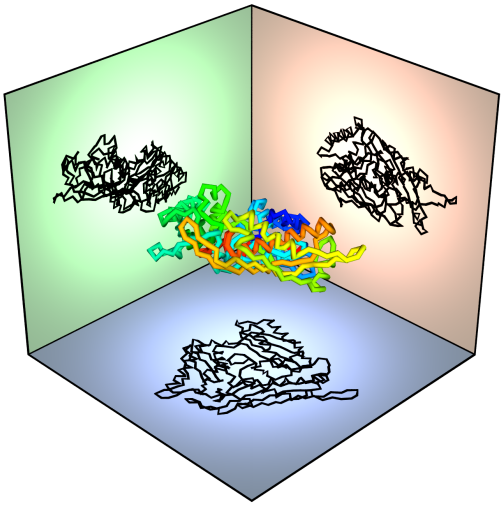


Figure 3.4: Orthogonal projections of a protein backbone. Clearly we will want virtual knot invariants to evaluate which virtual knots form on virtual closure of the projections.

3.2.1 Sphere closure methodology

The step-by-step procedure for analysing the knotting of an open curve using sphere closure is as follows:

1. Generate a set of 100 uniformly distributed points on the sphere, rotating this set if required to ensure that one point lies on the positive z axis [120].
2. Take a projection of the curve by considering only its (x, y)-coordinates.
3. Find the crossings of this projection and determine which strand is over-crossing by referring to the z-coordinates.
4. Connect the ends together with a straight line, adding over-crossings at any intersections.
5. Find the Gauss code of the resulting closed knot diagram.
6. Simplify this Gauss code using Reidemeister moves I and II as much as possible.
7. Calculate the Alexander polynomial of this knot diagram.
8. Rotate the curve and the set of points on the sphere such that a new point aligns with the positive z axis.
9. Repeat steps 2-7 until each of the 100 points has been considered.

This procedure is equivalent to joining the ends with straight lines to points on a sphere of infinite radius. The 100 uniformly-distributed points on the sphere are generated according to the algorithm of [120], which is based on the minimum energy spacing of charged particles in a conducting sphere. We will explore this more in Sec 3.2.4

Used naively, this procedure can fail if three strands in projection intersect at the same point. In practice, this is very rare for most open curves that we will consider, and can often be rectified by a slight rotation of the points on the sphere.

3.2.2 Virtual closure methodology

We perform virtual closure as follows:

1. Generate a set of 100 uniformly-distributed points on the sphere, rotating this set if required to ensure that one point lies on the positive z axis [120].
2. Take a projection of the curve by considering only its (x, y)-coordinates.
3. Find the crossings of this projection and determine which strand is over-crossing by referring to the z-coordinates.
4. Find the Gauss code of the open knot diagram.

5. Simplify using Reidemeister moves I and II as much as possible, taking care not to unthread the ends.
6. Calculate the generalised Alexander polynomial of the resulting Gauss code.
   a) If the generalised Alexander polynomial indicates $v2_1$ or the other virtual knots which share the same invariant value, notably $v3_2$, calculate the Jones polynomial of the open diagram to discriminate between them.
   b) If the generalised Alexander polynomial is zero, return to the projection and perform a closure with over-crossings as in sphere closure.
   c) Calculate the Alexander polynomial of the closed diagram.
   d) Return to the projection again and perform a closure using under-crossings.
   e) Calculate the Alexander polynomial of this new closed diagram.
   f) If the Alexander polynomials of these two diagrams differ, then the projection corresponds to a virtual knot.
7. Rotate the curve and the set of points on the sphere such that a new point aligns with the positive z axis.
8. Repeat steps 2-7 until each of the 100 points has been considered.

Here we say that if the over and under-closure of a projection produce different classical knots then the virtual closure produces a virtual knot. We do not currently have a proof for this, but for all the minimally genus one virtual knots we know of, replacing the virtual crossing(s) with over and under-crossings results in different classical knots.

3.2.3 Calculation of invariants

Here we will detail the methods we use to calculate each of the invariants used in sphere and virtual closure.

Alexander polynomial

The calculation of the Alexander polynomial, $\Delta(t)$, is most easily explained using knot diagrams, although in practice we use the Gauss code directly. The steps are as follows [3]:

1. Perform as many Reidemeister I and II moves as possible.
2. Choose an orientation for the knot and an arbitrary starting point on the knot diagram and number the crossings encountered when traversing

the diagram in order, exactly as is done in the Gauss code. Note also the crossing sign.

3. Number the arcs of the knot diagram, from under-crossing to under-crossing, in the same way.
4. Construct an $n \times n$ matrix, where n is the number of crossings in the diagram. The rows of this matrix correspond to crossings, and the columns to arcs.
5. Each crossing will have one incoming arc, one outgoing arc, and one over-crossing arc. For each matrix element (a, b), where a is the crossing in question, and b is the arc number, fill the element according to these rules:
   - For the incoming arc, enter -1.
   - For the outgoing arc, enter t if the crossing sign is positive and $1/t$ if it is negative.
   - For the over-crossing arc, enter $1 - t$ if the crossing sign is positive and $1 - 1/t$ if it is negative.
   - Every other element of the matrix is zero.
6. Calculate the determinant of any *minor* of the matrix. A minor is the matrix resulting from deleting a row and a column from the original matrix.

This final determinant is the Alexander polynomial. Depending on the minor chosen however, the polynomial may differ. By multiplying by $\pm t^m$, where m is any integer, all polynomials can be reproduced. Standard forms of the Alexander polynomial include the lowest power of t being a constant term, and the absolute value of the lowest negative power of t equalling the highest power of t. The latter form is possible because the Alexander polynomial may always be expressed in a symmetric form, where the coefficient of each positive power of t is equal to the coefficient of its partner negative power of t. For example, standard forms of the Alexander polynomial for the figure-eight knot are $\Delta(t) = t^2 - 3t + 1$ and $\Delta(t) = t - 3 + t^{-1}$.

Let's calculate the Alexander polynomial for the figure-eight knot by way of example. We use the diagram in Fig 3.5 to perform the calculation. The matrix we obtain from following the above rules is:

$$\begin{pmatrix} -1 & t & 1-t & 0 \\ 1/t & 1-1/t & 0 & -1 \\ 0 & -1 & 1/t & 1-1/t \\ 1-t & 0 & -1 & t \end{pmatrix}$$

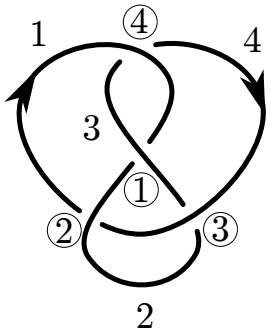


Figure 3.5: Arcs and crossings in the knot 4_1 labelled for calculation of the Alexander polynomial. Crossing numbers are circled and arc numbers are bare. Crossings 1 and 4 are positive, while 2 and 3 are negative.

We can then choose any minor we wish to calculate the determinant of. If we choose to delete row 1 and column 1 we obtain

$$\Delta(t) = \begin{vmatrix} 1-1/t & 0 & -1 \\ -1 & 1/t & 1-1/t \\ 0 & -1 & t \end{vmatrix} = 1 - 3t^{-1} + t^{-2}$$

We can then multiply this result by t to obtain the standard form $\Delta(t) = t - 3 + t^{-1}$.

For a complex knot diagram with many crossings, calculating the full symbolic polynomial can be very time consuming. We can substitute t for a constant without losing the invariant properties of the Alexander polynomial. However, as the polynomial is only invariant up to t^m, an arbitrary constant can produce an infinite number of valid answers. Using $t = 1$ avoids this problem, but $\Delta(1) = \pm 1$ for all knots. A standard substitution which maintains usefulness as an invariant is $t = -1$. As -1^m may equal 1 or -1, we take $|\Delta(-1)|$, giving us an invariant derived of the Alexander polynomial known as the determinant. The determinant has strictly less discriminating power than the full symbolic polynomial as many different polynomials can evaluate to the same constant. Of particular note, knots 4_1 and 5_1 both have determinant 5.

We can mitigate this loss of power by using more constants. Any constant for which $|t| = 1$ will work, which gives us the *roots of unity* $e^{2\pi ik/n}$, where k and n are integers. We use the roots $e^{2\pi i/3}$ and $e^{2\pi i/4} = i$ in addition to -1, which together form a surprisingly powerful invariant for simple knots. Of knots of 11 crossings or fewer, the simplest knot for which these constants are a poorer invariant than the symbolic polynomial is 9_4.

Generalised Alexander polynomial

The calculation of the generalised Alexander polynomial, $\Delta_g(s,t)$, bears some similarity to that of the Alexander polynomial in that a matrix is constructed based on crossing information and the determinant taken to produce the polynomial. However, the exact procedure differs in a number of ways, making the calculation more involved. Again, this is most easily explained using knot diagrams, although we use the Gauss code in practice. We follow the prescription given by Sawollek [117], which is as follows:

1. Take a virtual knot diagram, choose an orientation and number and sign the classical crossings as usual, performing Reidemeister moves where appropriate.
2. Associate with the n^{th} classical crossing a matrix M_n, which is M_+ if the crossing sign is + and M_- if the sign is −, where M_+ and M_- are defined as:

$$M_+ = \begin{pmatrix} 1-x & -y \\ -xy^{-1} & 0 \end{pmatrix} \quad \text{and} \quad M_- = \begin{pmatrix} 0 & -x^{-1}y \\ -y^{-1} & 1-x^{-1} \end{pmatrix}$$

3. Construct the $2n \times 2n$ block matrix, M, where each diagonal element is given by the matrix M_n and all other elements are zero.

$$M = \begin{pmatrix} M_1 & 0 & \cdots & 0 \\ 0 & M_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \cdots \\ 0 & 0 & \cdots & M_n \end{pmatrix}$$

4. Now divide the virtual knot diagram into edges which run from classical crossing to classical crossing. Note these are not arcs as before and so do not continue through over-crossings. They may intersect at virtual crossings.
5. At each crossing, two edges will begin and two edges will end. Label the beginnings and endings of each edge at crossing n according to Fig 3.6.
6. For each edge, assign the beginning label to its corresponding end label: $(i,a) \mapsto (j,b)$, where i and j are the beginning and ending crossing numbers respectively and a and b are the L and R labels. The beginning label will necessarily have a + superscript and the ending label a − superscript.
7. Construct the $2n \times 2n$ *permutation matrix, P*, associated with these assignments. The rows of this matrix correspond to labels $1^+_L, 1^+_R, 2^+_L, 2^+_R, ..., n^+_L, n^+_R,$, while the columns correspond to labels $1^-_L, 1^-_R, 2^-_L, 2^-_R, ..., n^-_L, n^-_R,$. For each

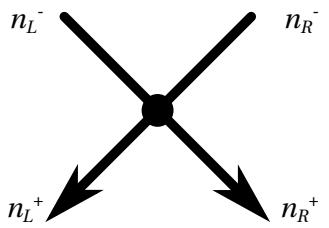


Figure 3.6: The edge labels at crossing n, used when calculating the generalised Alexander polynomial. The crossing here is any classical crossing. Which strand crosses over and which crosses under does not change the labels.

assignment of beginning and ending labels, enter a 1 in the corresponding matrix element. All other elements are zero.

8. Calculate $-1^{w(K)}det(M-P)$, where $w(K)$ is the writhe of the virtual knot diagram.

9. Perform the substitution $x = st$ and $y = -t$.

This final quantity is the generalised Alexander polynomial. The substitution is necessary to turn the form given by Sawollek [117] into that used by the virtual knot table [112]. The form in (x, y) is an invariant up to x^m, where m is an integer, so the final form is invariant up to factors of $(st)^m$.

In comparison to the Alexander polynomial, it is clear that this calculation is more demanding, requiring the calculation of the determinant of a $2n \times 2n$ matrix as opposed to a $(n-1) \times (n-1)$ matrix. While Reidemeister moves can ease this for curves which may have complex initial projections but which are topologically simple, this matrix size difference cannot be avoided.

As an example, we shall calculate the generalised Alexander polynomial of $v2_1$ as shown in Fig 3.7. Included in the figure are the edge labels at each crossing we will need. Both crossings in this instance are positive, thus:

$$M = \begin{pmatrix} 1-x & -y & 0 & 0 \\ -xy^{-1} & 0 & 0 & 0 \\ 0 & 0 & 1-x & -y \\ 0 & 0 & -xy^{-1} & 0 \end{pmatrix}$$

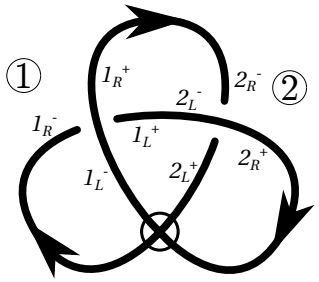


Figure 3.7: Edges around crossings in the knot $v2_1$ labelled for calculation of the generalised Alexander polynomial. Crossing numbers are circled.

The permutation matrix, P following the labels in Fig 3.7 is:

$$P = \begin{array}{c} \\ 1_L^+ \\ 1_R^+ \\ 2_L^+ \\ 2_R^+ \end{array} \begin{array}{c} \begin{array}{cccc} 1_L^- & 1_R^- & 2_L^- & 2_R^- \end{array} \\ \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix} \end{array}$$

Calculating $det(M-P)$ gives us $x^2y^{-1}+x^2+xy-xy^{-1}-y-1$. The writhe is two, so $-1^{w(K)}=1$. Substituting $x=st$ and $y=-t$ gives the final result:

$$\Delta_g(s,t) = (s^2-s)t^2+(-s^2+1)t+s-1$$

Again, the full symbolic calculation of the generalised Alexander can take a long time. By sacrificing some discriminating power, we can substitute constants as we did for the Alexander polynomial to produce a fast to calculate invariant. If we take $|\Delta(s,t)|$, and require that $|st|=1$ analogously to the procedure for the Alexander polynomial, we find that s and t must be roots of unity as before. For simple virtual knots, such as those minimally genus one virtual knots we are aware of up to four crossings, the combinations $(s=-1, t=e^{2\pi i/3})$, $(s=-1, t=i)$, and $(s=e^{2\pi i/3}, t=i)$ are as good as the symbolic generalised Alexander. Unfortunately, even the full generalised Alexander does not distinguish $v2_1$ and $v3_2$, the two simplest minimally genus 1 virtual knots. To discriminate between these two, we use the Jones polynomial.

Jones polynomial

The Jones polynomial calculation differs from the Alexander-type polynomials discussed so far. Instead of constructing a matrix from crossing information

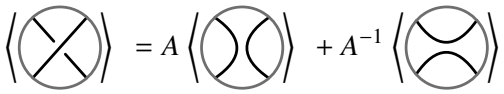


Figure 3.8: The skein relation used in calculating the Kauffman bracket variant of the Jones polynomial.

and calculating a determinant, the knot diagram itself is manipulated according to a certain algebra until the final polynomial is reached [1]. The form of the Jones polynomial we use is derived from the *Kauffman bracket* and is based on the relation shown in Fig 3.8. What is shown is a procedure, called a *skein relation*, applied to a single crossing of a knot diagram, which results in two new, smoothed diagrams. Also associated to each new diagram is a factor of A or A^{-1}. By successively applying this relation to all crossings in a knot diagram, a collection of unknots will eventually result. A diagram with a single unknot is given a value of 1. Each additional unknot in a diagram is substituted for a factor of $(-A^2 - A^{-2})$. The polynomial resulting from this process is the bracket polynomial, or Kauffman bracket. The bracket polynomial however is not invariant under Reidemeister move one. This can be rectified by multiplying the polynomial by $(-A^3)^{-w(K)}$ where $w(K)$ is the writhe of the knot, K, being considered. To complete the calculation, factors of A are substituted by $q^{-1/4}$ which is then identical to the Jones polynomial. An example calculation for the virtual knot $v2_1$ is shown in Fig 3.9 and the more complex classical knot 4_1 in Fig 3.10.

What can be seen immediately is that if each crossing is split into two diagrams, and each of those diagrams must be split into two more diagrams and so on, the final number of diagrams is going to grow exponentially with crossing number. This makes the Jones polynomial the most computationally expensive invariant to calculate of those described so far. The Mathematica package KnotTheory` implements an optimisation of the Jones which attempts to maintain a 'computation front', wherein neighbouring crossings are smoothed in turn and any unknots are tidied up during the calculation. This seeks to reduce the number and complexity of diagrams stored at one time by simplifying as much and as often as possible. Note that the simplifications here do not include

[1] In fact, the Alexander polynomial may be calculated in a similar manner, sometimes referred to as the Alexander-Conway polynomial. This is not the most practical computationally however, which is why we instead use the algorithm already described.

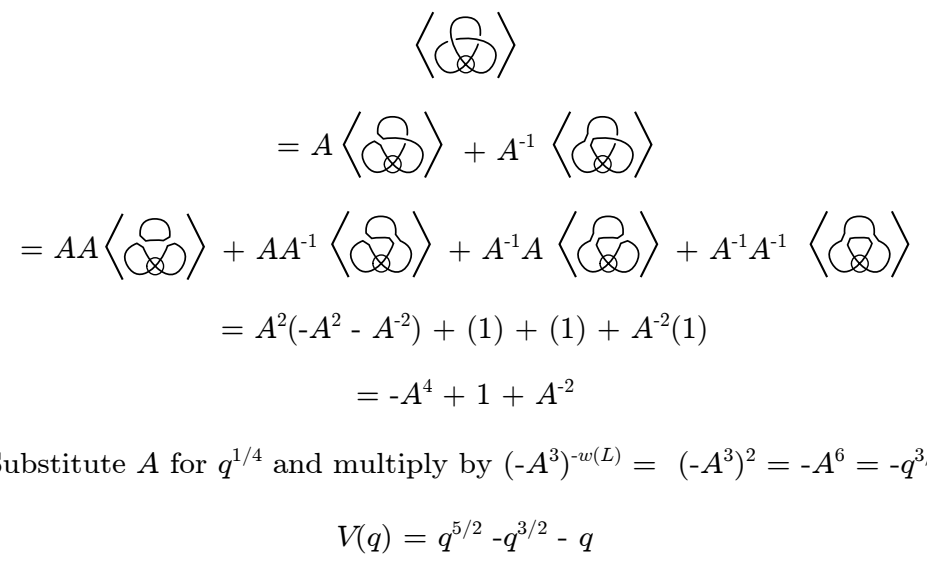


Figure 3.9: Calculation of the Jones polynomial of the $v2_1$ virtual knot using the Kauffman bracket.

performing Reidemeister moves on the smoothed diagrams as this affects the final polynomial but instead means removing isolated loops as they appear and absorbing them into the coefficients of that diagram. We implement this optimisation also.

We encode knot diagrams used in the Jones calculation using the *planar diagram* presentation. In planar diagram presentation, an oriented knot diagram is split into edges, running from one classical crossing to the next, and each edge is numbered in order. Each crossing then is given a label X_{ijkl} where i is the incoming under-crossing edge and j, k and l are the remaining edges encountered counter-clockwise from i. The whole diagram is then encoded as a list of these X labels. Fig 3.11 shows a labelled 4_1 knot and gives its planar diagram presentation in the caption.

The Kauffman bracket in planar diagram language then, involves the removal of one X and a subsequent relabelling of the edges which are now connected. Taking our cue from the KnotTheory` package again, it is helpful here to introduce some new notation to keep track of the joined edges. We introduce a *point* P_{ij}, which is a point which connects to edges i and j after smoothing. The order of i and j here does not matter, unlike at a crossing. The Kauffman skein relation in planar diagram notation then becomes $\langle X_{ijkl}\rangle = A\langle P_{il}P_{jk}\rangle + A^{-1}\langle P_{ij}P_{kl}\rangle$. We recognise that $P_{ab}P_{bc}$ is an edge which begins at a, ends at c and contains b which cannot now participate in crossings, hence we make the substitution

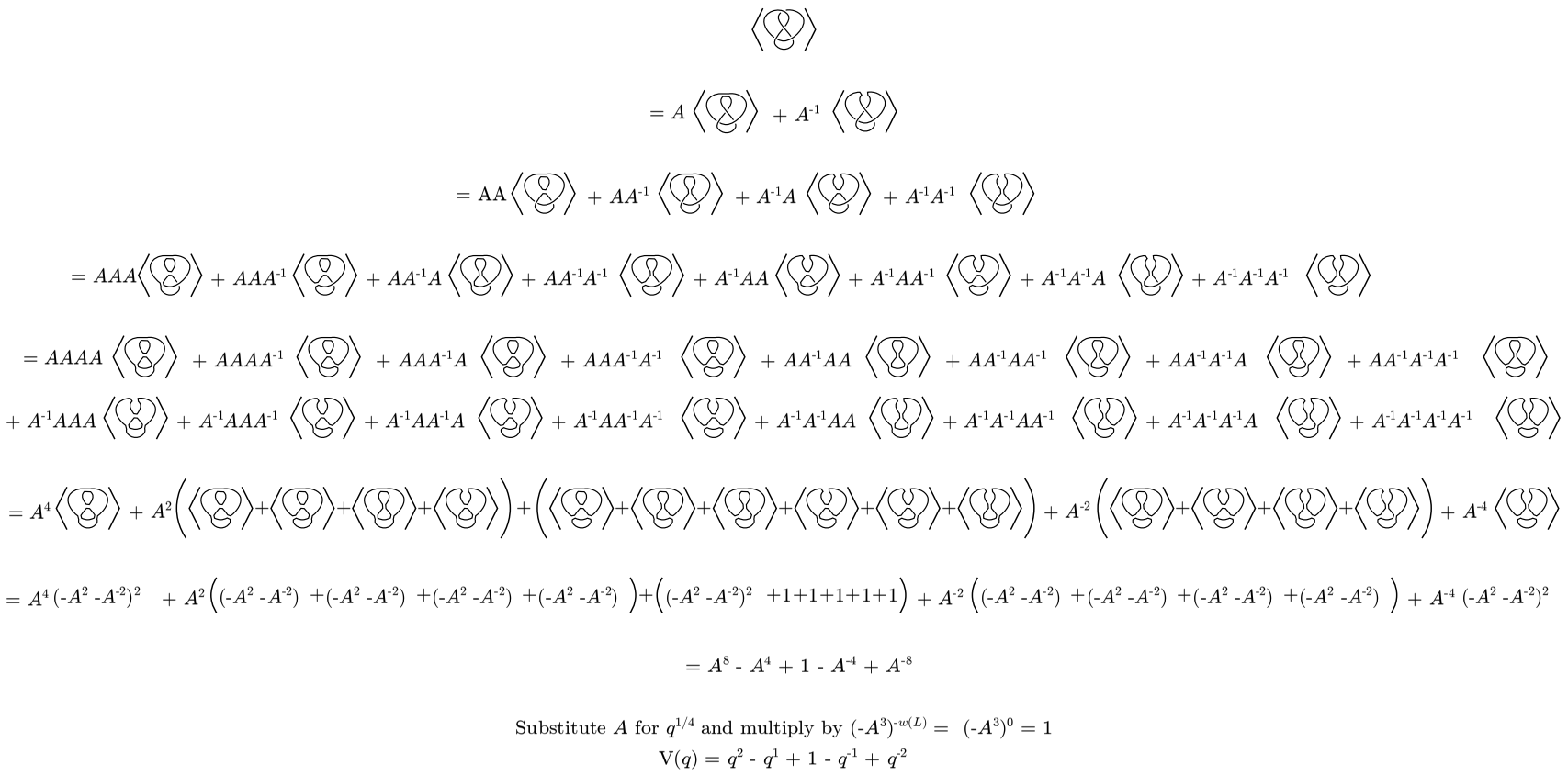


Figure 3.10: Calculation of the Jones polynomial of the 4_1 knot using the Kauffman bracket.

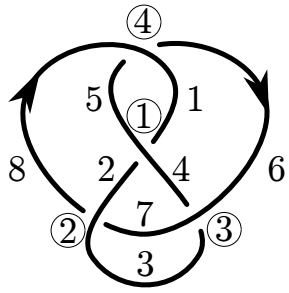


Figure 3.11: The 4_1 knot with edges and crossings numbered, with crossings circled. The planar diagram label for each crossing in order is $X_{1524}X_{7283}X_{3647}X_{5168}$.

$P_{ab}P_{bc} = P_{ac}$. Further we recognise that $P_{ab}P_{ab} = P_{aa}$, which represents an isolated loop. The implementation of the Jones in planar diagram notation then involves smoothing all crossings, applying the rules to simplify points until all that is left are isolated loops and substituting $(-A^2 - A^{-2})$ for each P_{aa}. In the optimised routine all possible simplifications and substitutions are done at each step. The entire polynomial can then be divided by $(-A^2 - A^{-2})$ to give the bracket polynomial, before multiplying by $(-A^3)^{-w(K)}$ and substituting $q^{-1/4}$ for A as usual.

As we only calculate the Jones in order to distinguish between $v2_1$ and $v3_2$, we do not need the power of the full symbolic Jones polynomial. By substituting a constant here again, we can speed up the calculation while retaining the properties we want. Unlike the Alexander and generalised Alexander, the Jones polynomial is not invariant up to factors of q or similar, so we can in fact substitute any constant we like. We use $|V(q = -1)|$, which is enough to distinguish $v2_1$ and $v3_2$. By taking the absolute value, we lose the power to detect chirality, but we are not going to focus on chirality or any other virtual knot mirrors in the analysis in this thesis. The invariant values for a selection of classical and virtual knots are given in Table 3.1.

3.2.4 Number of closure directions necessary

When performing closure analyses, an obvious question is how many closure directions are necessary? The answer depends on how much detail one wants, and the complexity of the curve being analysed. It will take fewer closures if all one wants to know is the most dominant knot, compared to identifying the five most common knots in order with accurate fractions of each. In most of the

Knot	$\lvert\Delta(-1)\rvert$	$\lvert\Delta(e^{2\pi i/3})\rvert$	$\lvert\Delta(i)\rvert$	$\lvert V(-1)\rvert$
0_1	1	1	1	1
3_1	3	2	1	3
4_1	5	4	3	5
5_1	5	1	1	5
5_2	7	5	3	6
6_1	9	7	5	9

Knot	$\lvert\Delta_g(-1,e^{2\pi i/3})\rvert$	$\lvert\Delta_g(-1,i)\rvert$	$\lvert\Delta_g(e^{2\pi i/3},i)\rvert$	$\lvert V(-1)\rvert$
$v2_1$	3	4	5	2
$v3_2$	3	4	5	4
$v4_{12}$	3	8	5	5
$v4_{36}$	3	8	9	4
$v4_{37}$	0	4	8	2
$v4_{43}$	7	8	9	5
$v4_{64}$	3	4	0	4
$v4_{65}$	3	8	9	5
$v4_{94}$	3	4	5	5
$v4_{100}$	3	0	8	4

Knot	$\Delta_g(s,t)$
$v2_1$	$s^2+s^2/t+st-s/t-t-1$
$v3_2$	$s+s/t+t-1/t-t/s-1/s$
$v4_{12}$	$s^2/t+s^2/t^2-st-s/t-2s/t^2-t^2+t-1/t+$ $1/t^2+2t^2/s+t/s+1/st-t^2/s^2-t/s^2$
$v4_{36}$	$t-1-1/t+1/t^2+t^2/s-2t/s+2/st-1/st^2-$ $t^2/s^2+t/s^2+1/s^2-1/s^2t$
$v4_{37}$	$s^4-s^4/t^2+s^3t+s^3/t^2-s^2t+s^2/t-st^2-s/t+t^2-1$
$v4_{43}$	$s^3+s^3/t+s^2t-s^2/t-st-s$
$v4_{64}$	$s^2/t+s^2/t^2-s/t-s/t^2+t^2/s+t/s-t^2/s^2-t/s^2$
$v4_{65}$	$-s^2t+s^2+s^2/t-s^2/t^2-st^2+2st-$ $2s/t+s/t^2+t^2-t-1+1/t$
$v4_{94}$	$s^3+s^3/t+s^2t-s^2/t-st-s$
$v4_{100}$	$s^4+s^4/t+s^3t-s^3/t-s^2t+s^2/t+st-s/t-t-1$

Table 3.1: Table of numerical knot invariants for a number of classical and virtual knots. Included are $|\Delta(t)|$, the Alexander polynomial at the numerical values used, $|\Delta_g(s,t)|$ the generalised Alexander polynomial, both at the numerical values used and the full symbolic expression, and $|V(-1)|$ the Jones polynomial with $q=-1$. As the generalised Alexander polynomial of all classical knots is zero, these columns are omitted. Virtual knots have no Alexander polynomial and so these columns are omitted. In the cases where chiral mirrors give different knots, only one mirror is given.

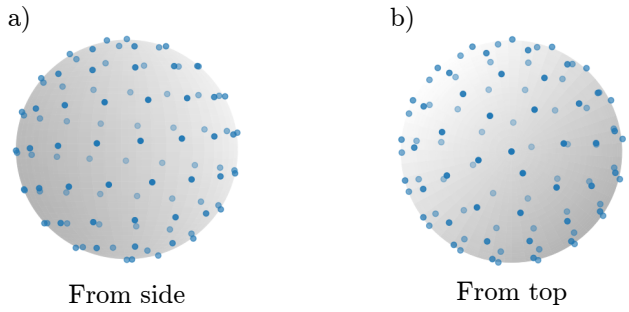


Figure 3.12: Distribution of 100 approximately uniform points on the sphere according to the algorithm of [120]. These give the projection directions we use for our closure analyses.

analysis in this thesis, we are interested in accurately determining the fraction of closures which give the most common knot, as well as the fraction of closures which result in the three categories of unknots, classical knots, and virtual knots.

First, to obtain the most accurate results, we want to choose directions on the sphere as uniformly as possible. We use the generalised spiral points described by Rakhmanov et al. [120] to determine these directions. This is a fast algorithm which tries to capture the positions electrons in a conducting sphere will assume to minimise their energy. Fig 3.12 shows the results of this algorithm for 100 points, which is the number of projections we use.

The use of 100 closure directions was initially taken from the convention used by KnotProt [23], but to show that this is sufficient using the point choosing algorithm described, see Figs 3.13 and 3.14. Each of these figures show how the fraction of each knot type varies as more closures are chosen, with the a) figures using sphere closure and the b) figures using virtual closure. Fig 3.13 analyses the knotting of a simple open trefoil curve. The knotting spectrum for this curve is not complicated under either sphere or virtual closure, containing two major components and two trace components in each closure scheme. The black vertical line marks the 100 closures point and while increasing the number of closures reduces the noise, the analysis does not change a great deal.

Fig 3.14 analyses instead the knotting of the protein with PDB ID 4XIX, chain A [121]. The exact details of this protein are not important right now other than the fact that it is conformationally much more complex than the open trefoil above and will be more typical of the curves analysed later in this thesis. The knotting spectrum here is more involved, particularly under virtual closure where at least ten different knot types are present. The noise at 100 closures

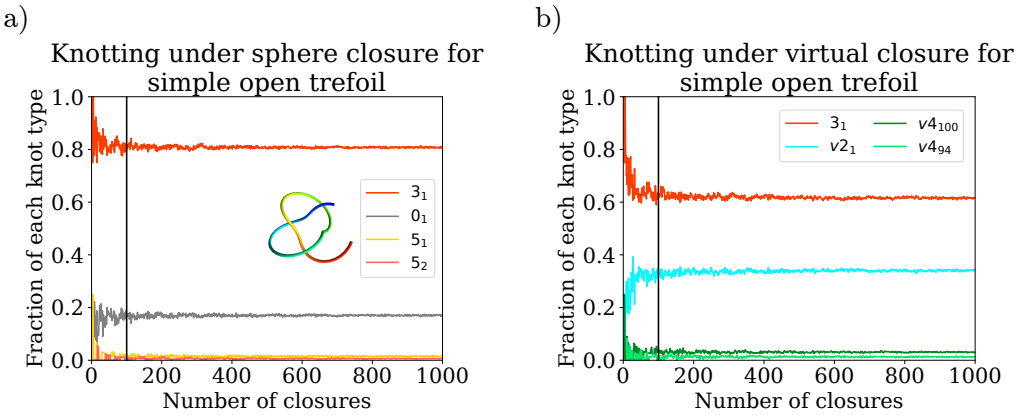


Figure 3.13: How the fraction of a given knot type in closure analyses vary for a simple open trefoil (inset). Sphere closure is used in a) and virtual closure in b). The black vertical line marks 100 closures.

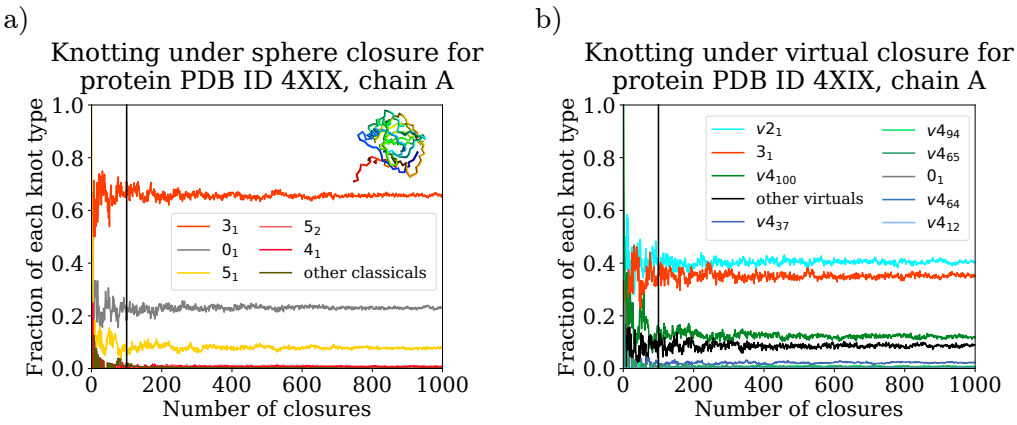


Figure 3.14: How the fraction of a given knot type in closure analyses vary for the protein with PDB ID 4XIX, chain A (inset). Sphere closure is used in a) and virtual closure in b). The black vertical line marks 100 closures. We do not distinguish knot types beyond those given in the legends, and so closures which produce more complex knots are gathered together under other classicals and other virtuals.

is greater here than for the simple curve, but again not much changes beyond this point. The order of the knots remains stable and at very close to the same fraction.

If one is only interested in the knotting of one specific curve, taking as many as 1000 closures would ensure the results are accurate for curves of the complexity of proteins. Usually in this thesis however we will be most interested in the knot statistics of ensembles of curves and increasing the number of closures used by so much would make the computational time unreasonably long.

3.3 Initial exploration of virtual closure

Having covered the algorithmic details of virtual closure, here we will discuss what virtual closure can tell us about open curve conformations. This will include the different categories we might expect the results to fall into, and some insights into the relation between the knots seen from related projection directions. This will point towards an understanding of the shape of knot space and the place of virtual knots within it which we have not seen highlighted elsewhere. Also covered are possible extensions to the analysis of slipknots using virtual knots.

3.3.1 Strong and weak knotting

Before considering the spatial aspects of a closure analysis, we will look first just at the overall fractions of each knot type and the different possible distributions we can see. Taking sphere closure first, the analysis of a space curve will in general return fractions of different classical knot types as well as some fraction of unknots. There are a number of ways we can categorise analyses like these. We could take the single most common knot as representative of the knotting of the curve, but this may lead to situations such as the unknot covering 40% of projections being taken as representative, despite the fact that knotting is more common than unknotting. To avoid this, we say that a curve is *knotted* under sphere closure if 50% or fewer closures are unknots. Conversely, a curve is *unknotted* under sphere closure if unknots make up greater than 50% of closures.

For knotted curves, we make a further distinction. If a single classical knot type covers 50% or more closures, then that knot type may be taken as representative of the curve. We call curves where this is the case *strongly classically knotted*. If, for example, the trefoil is the knot that dominates in such a curve, the curve would be strongly trefoil knotted. This leaves a final category where the curve is knotted, but no single classical knot type covers a majority of closures. We call these curves *weakly classically knotted*. With this, we have three broad categories for the results of a sphere closure analysis: unknotted, strongly classically knotted and weakly classically knotted.

As ever, with the introduction of virtual knots this classification becomes more complicated. Under virtual closure, we maintain the same definition of unknotting, where more than 50% of virtual closures are unknotted. Knotted

curves now have more possibilities however. For strong knotting, we distinguish between curves where a classical knot dominates and where a virtual knot dominates. Strong classical knotting remains, where a single classical knot type covers 50% or more closures, but we add the *strong virtual knotting* category, where a single virtual knot type covers 50% or more closures. We have not seen a curve where both of these criteria are fulfilled simultaneously.

We make further distinctions for weak knotting. Curves where 50% or more closures give classical knots, with no single type as dominant, we call weakly classically knotted as before. If instead virtual knots cover 50% or more closures with no single dominant type, we call the curve *weakly virtually knotted.* Finally, curves where classical knots taken together with virtual knots account for 50% or more closures, but where neither category alone does this, we call *weakly totally knotted.* So for virtual closure the categories we define are unknotted, strongly classically knotted, strongly virtually knotted, weakly classically knotted, weakly virtually knotted and weakly totally knotted.

When comparing the same curve analysed with each method, many closure directions may return the same knot type (including the unknot) on virtual closure as for sphere closure. Depending on the particular conformation of the curve, there may in fact be very little difference between the two methods. Where there is a difference, the only possibilities are classically knotted closures becoming virtual knots, and unknotted closures becoming virtual knots. It is not possible that a classical knot becomes an unknot, and so every knotted curve under sphere closure is also a knotted curve under virtual closure. It is very possible that a strongly knotted curve may be reclassified as weak knotted however.

One could argue that the use of 50% as the cut-off for knottedness is arbitrary. In its defence, 50% is the boundary at which a knot type or category of knot types achieves a majority over the other types. There is nothing to stop one from requiring a higher or lower threshold for knottedness however, if that is of use to the researcher. Care would have to be taken when using a lower threshold to determine between different weak and strong knotting categories but this will not be explored further here. In our results we will present data on the distribution of the coverage of the most common knot type in different systems, and from these one can see how knotting classifications would vary by requiring a different cut-off.

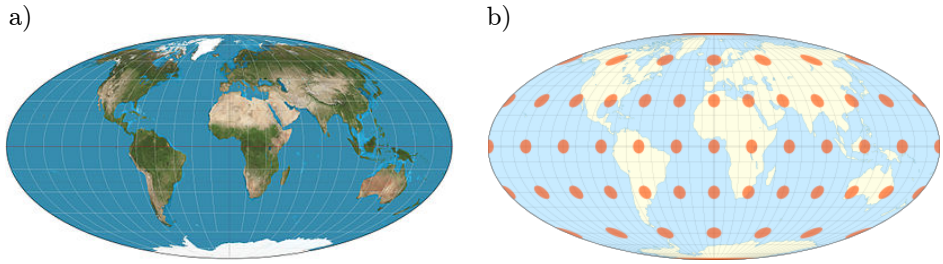


Figure 3.15: Mollweide projections of the Earth, taken from [122]. a) shows a standard map view, and b) indicates the distortions applied by the projection. All the red shapes in b) are circles on a globe. Crucially, the Mollweide projection preserves the area of these circles, although they become more distorted towards the poles and further from the prime meridian.

3.3.2 Knot globes

More structural detail can be obtained by considering the closure directions in addition to the overall knot fractions. For a given curve, a map of knot type with closure direction can be constructed. This can be visualised as a sphere with regions coloured according to knot type obtained when closing to/projecting from that region. We call such visualisations, *knot globes*.

In displaying the knot globes in this thesis, in addition to showing a single view of each globe as a sphere, we will plot 2D projections. In particular we will be using the *Mollweide projection* as it preserves area and we are interested in the areas covered by each knot type. Fig 3.15 shows the Earth plotted with a Mollweide projection as a way of familiarising the reader with the shape distortions that accompany this projection.

Fig 3.16 shows knot globes for the protein with PDB ID 4K0B, chain A [123], one using sphere closure and the other using virtual closure, as well as area preserving maps of each. The protein is kept in the same orientation in each for comparison. The borders between knot types happen when, as the protein is rotated, an end-point has passed a strand in projection, either adding or removing crossings in the projected diagram, resulting in a change of knot type. It can be seen that borders between knot types under sphere closure remain borders under virtual closure, and there are new borders under virtual closure also, picked out by the wider variety of virtual knot types.

Closer inspection shows us that with virtual closure, between regions of classical knotting there is almost always a region of virtual knotting. To explain this we note that to change between classical knot types, an over-crossing must become an under-crossing in a (potentially non-minimal) knot diagram. If an

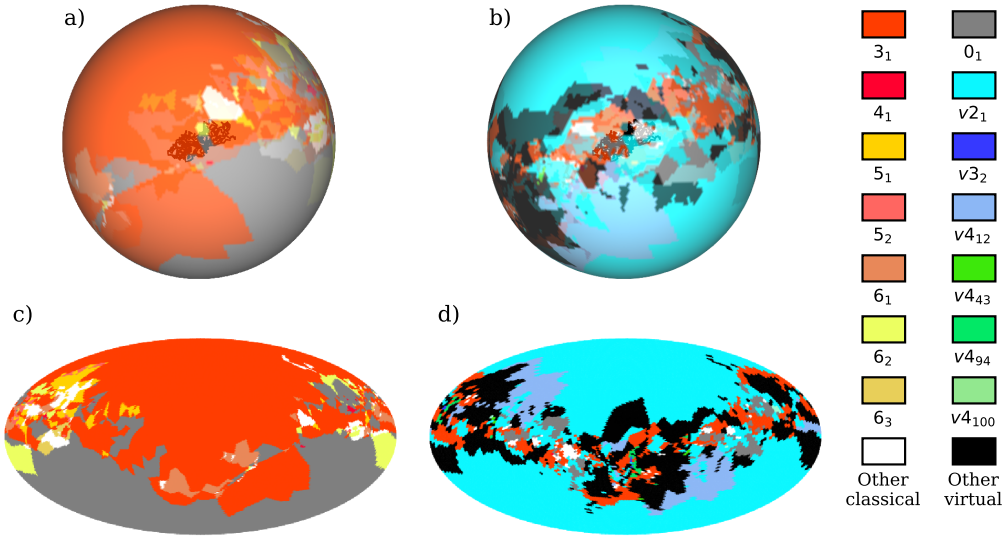


Figure 3.16: Knot globes for protein with PDB ID 4K0B, chain A using a) sphere closure and b) virtual closure. The globes here are translucent, allowing the protein to be seen, as well as the back of the globes. Shown in c) and d) are Mollweide maps of the globes a) and b) respectively.

open knot diagram is virtually closed to give a classical knot (or unknot), there can be no virtual crossings after virtual Reidemeister moves. To change an over-crossing to under-crossing or vice versa, an end-point must be moved past a strand to remove the crossing, and then either it or the other end-point pass a strand in the opposite orientation, adding a crossing. If the diagram is virtually closed throughout this process, the crossing will transition from over to virtual to under, or vice versa. This process is shown in Fig 3.17. If this crossing flip results in a knot type change then the crossing is topologically important and replacing it with a virtual crossing will produce a virtual knot. Thus different classical knot types on the knot globe tend to have a region of virtual knotting between them.

The exceptions to this rule come from very specific open curve conformations. For classical knot types to border each other under virtual closure, both ends must simultaneously cross a strand as the curve is rotated. This process is illustrated in Fig 3.18. As can be seen, at no point is a virtual crossing added. An extended border between classical knot types can be created as the end-points slide 'up and down' the strand shown in Fig 3.18 b). While this situation is uncommon in curves like proteins, it is more prevalent in random walks on lattices.

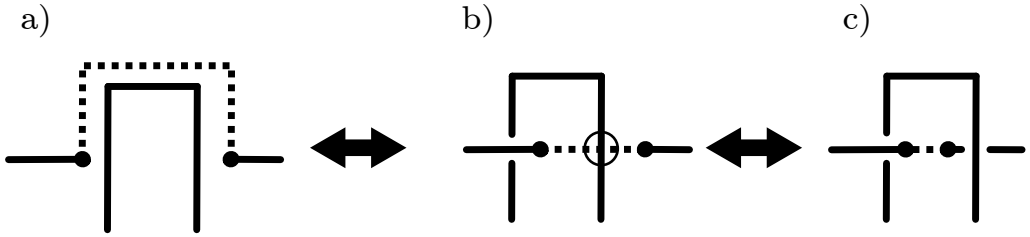


Figure 3.17: The flipping of a classical crossing as may happen by changing projection direction gradually. Closure path is shown by the dashed line. a) shows an initial scenario where no crossings need be added to join the ends. A linear sphere closure of this diagram would introduce two over-crossings which could then be removed by Reidemeister move II In b) the projection has changed such that one of the ends has introduced an irremovable over-crossing. A sphere closure here would introduce another over-crossing, allowing the diagram to be manipulated to resemble a). The virtual closure of this diagram however induces a virtual crossing as shown. In c), the other end-point has passed the strand but this time as an under-crossing. If this procedure involved a change in classical knot type between a) and c) then the knot type in b) must be a virtual knot.

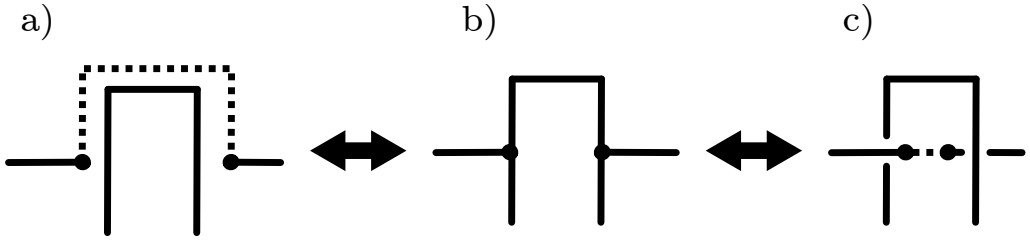


Figure 3.18: A potential transition between two classical knots under virtual closure. Closure path is shown by the dashed line. a) shows the initial state where a closure path may be drawn between the ends without inducing crossings. In b) the projection has changed such that the ends now both lie on a strand. This intermediate state exists on the boundary between knot types on the knot globe and does not produce a valid knot. In c), both end are now within the original loop and can be joined without inducing crossings. Two new classical crossings have been created and so this represents a transition between two classical knots under virtual closure.

Looking elsewhere shows us that when antipodal points, those on opposite sides of the sphere, give different knot types under classical closure, the virtual closure will identify both those points with a virtual knot. Thinking about this diagrammatically, if one projects a curve to obtain a knot diagram and closes the ends first with all over-closures and then all under-closures, this is equivalent to closing to antipodal points. As we conjectured earlier in Sec 3.2.2, when these two closures give different knots, a virtual closure should give a virtual knot. This need not be the same virtual knot, if the vertical plus horizontal mirrors of the virtual knot are not equivalent. In this way, the virtual closure globe can be constructed from the sphere closure globe, by matching antipodal points and determining which virtual knot lies between the different classical knot types.

To appreciate the variety of knottings possible in simple curves, we present a selection of knot globes for some random walks on a $6 \times 6 \times 6$ cubic lattice. The details of how we create this will be discussed in Chapter 5. The curve in Fig 3.19 is strongly classically knotted under virtual closure, whereas strong virtual knots are presented in Fig 3.20. Weak classical knotting under virtual closure is displayed by the curve in Fig 3.21, weak virtual knotting in Fig 3.22 and weak total knotting in Fig 3.23. As these curves are on lattice, situations where classical knots border classical knots under virtual closure, as in Fig 3.18, are more common than in the proteins example. Additional examples as well as graphs of the connected areas are given in Appendix A.

3.3.3 Knot space

As has been touched on in previous sections, the relation between classical and virtual knots is important when considering virtual closure. By flipping crossings in classical knots, the knot type can, of course, change. Fig 3.24 indicates which classical knots are related by crossing flips, for prime knots up to seven crossings. Virtual knots enter this picture by adding a middle step to a crossing flip, where the crossing is first changed to a virtual crossing, and then the opposite classical crossing. Along every directed edge in Fig 3.24, there exists a virtual knot where the flipped crossing is replaced with a virtual crossing and Table 3.2 gives their knot invariants. This gives an indication as to the shape of *knot space*.

All of these virtual knots are minimally genus one and so can occur on virtual closure, but not all minimally genus one knots lie along directed edges like these. As only one virtual crossing exists along each directed edge, virtual knots like

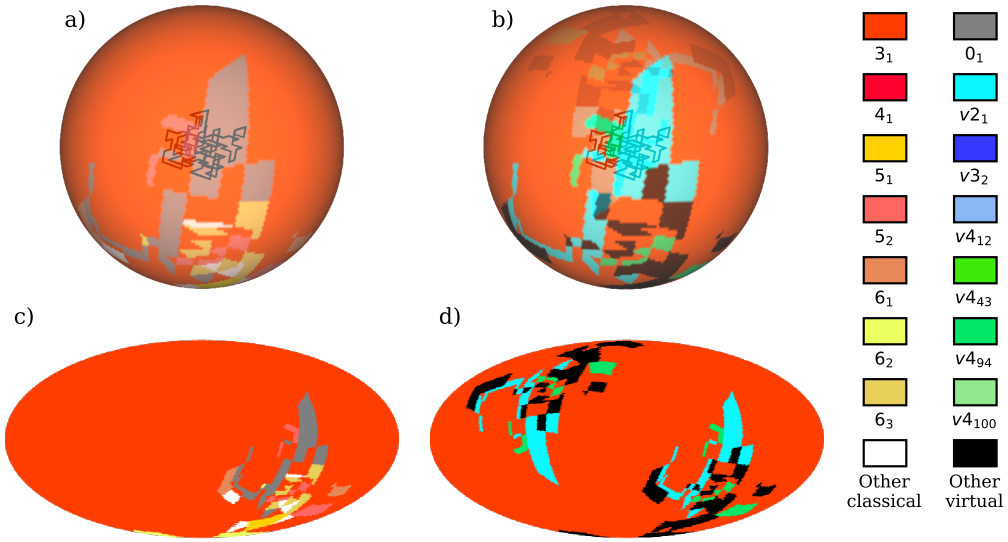


Figure 3.19: This self-avoiding lattice walk is strongly trefoil knotted under both sphere and virtual closure. See Fig 3.16 for other details.

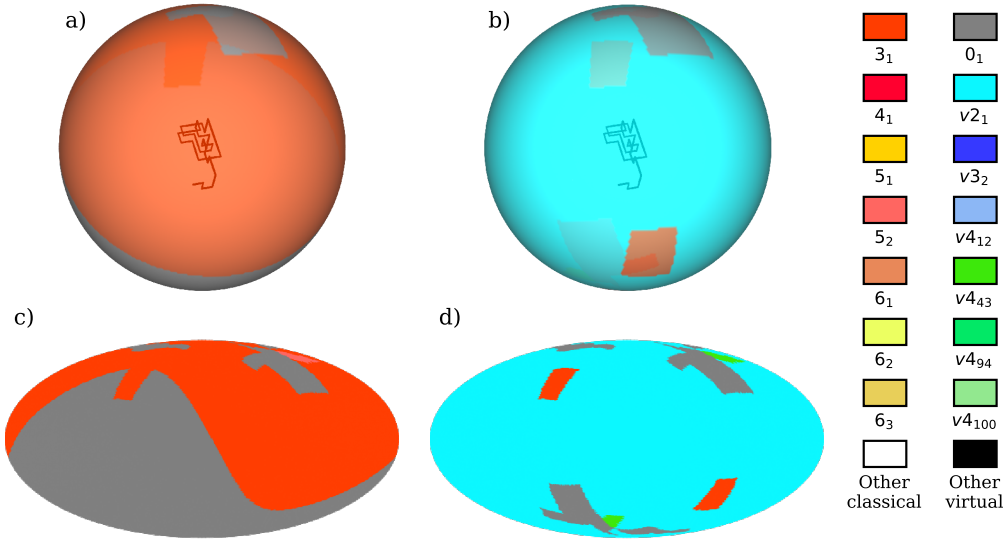


Figure 3.20: This self-avoiding lattice walk is unknotted under sphere closure, with a 52% coverage of the unknot, but strongly $v2_1$ knotted under virtual closure. See Fig 3.16 for other details.

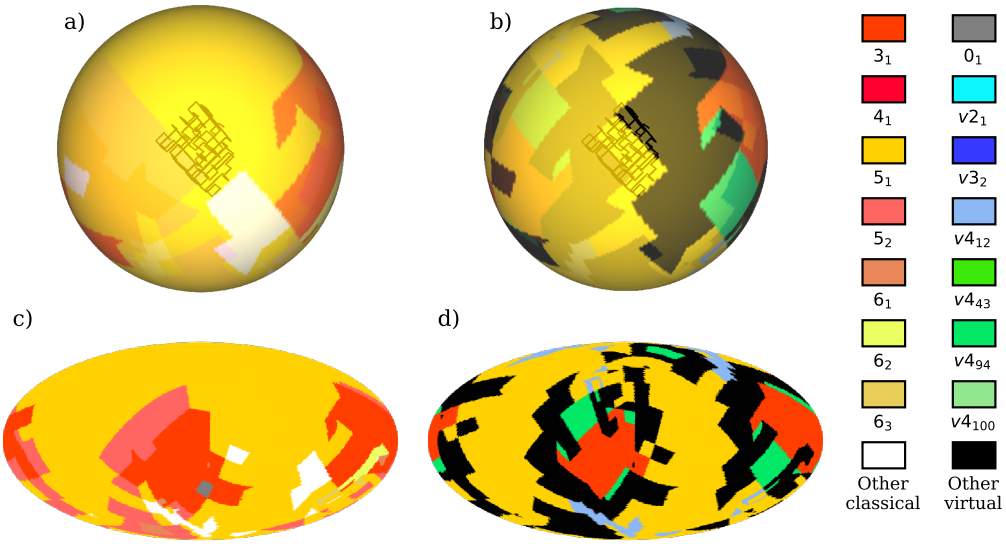


Figure 3.21: This self-avoiding lattice walk is strongly 5_1 knotted under sphere closure, but weakly classical knotted under virtual closure. The 5_1 areas under sphere closure have been eaten away by virtual knots under virtual closure, reducing its dominance to below 50%. See Fig 3.16 for other details.

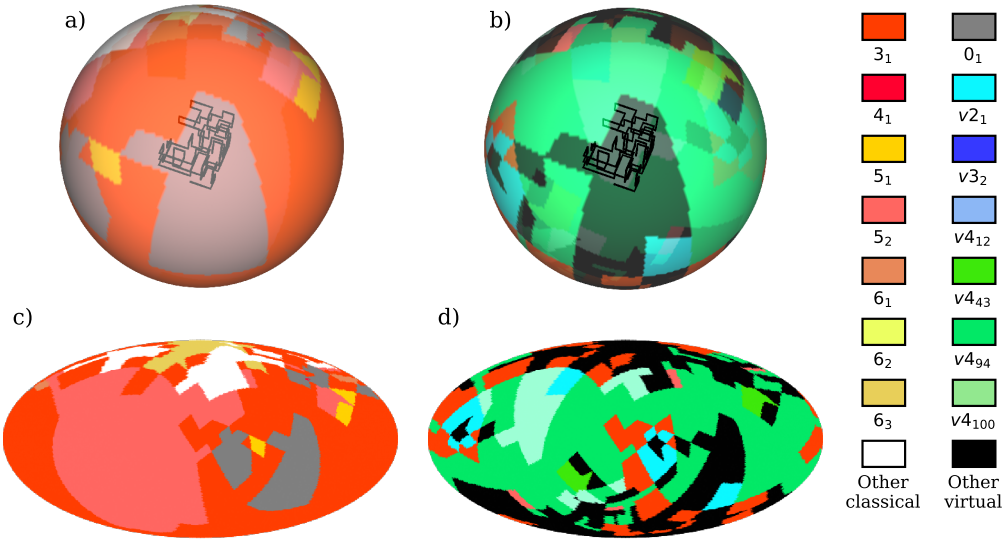


Figure 3.22: This self-avoiding lattice walk is weakly classically knotted under sphere closure, and weakly virtual knotted under virtual closure. No unknotted regions remain under virtual closure. See Fig 3.16 for other details.

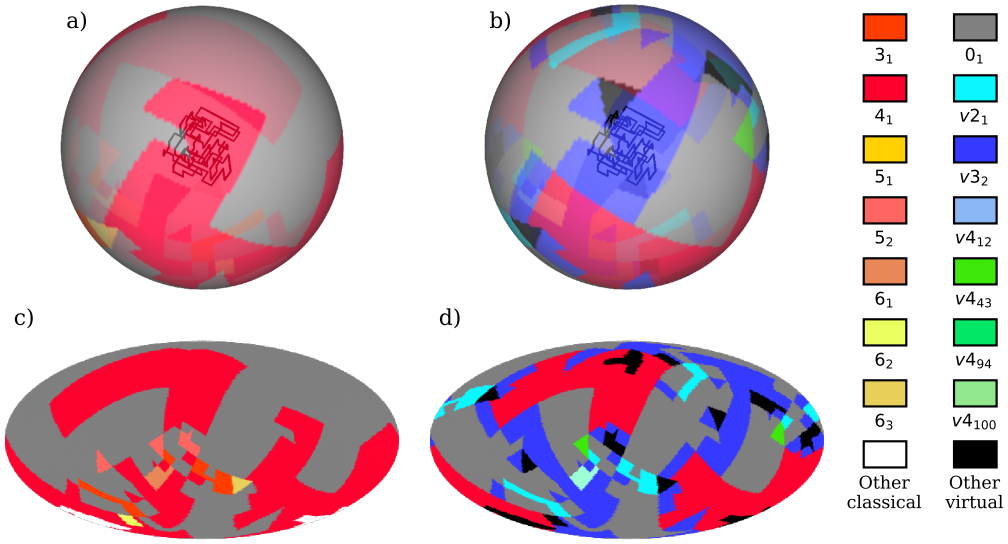


Figure 3.23: An unknotted self-avoiding lattice walk under sphere closure (52% unknot) which is weak total knotted under virtual closure. Unknots still make up 40% of virtual closures, with classicals comprising 25% and virtuals the remaining 35%. See Fig 3.16 for other details.

$v4_{94}$ which have two virtual crossings exist outside of this diagram. One could branch the directed edges to indicate more virtual crossings being added before crossings are fully flipped and this will capture a greater variety of virtual knots, including non-minimally genus one virtual knots. Indeed, by including non-minimal classical knot diagrams, every virtual knot could be reached in this way.

Fig 3.24 is not intended to be an exhaustive description of knot space, particularly as it only considers minimal diagrams. There are many features in this figure that appear significant but which are a product of the limited information presented. For example, any crossing flip in this figure always changes the minimal crossing number by at least two. This is due to the fact that all knots below eight crossings are *alternating* knots i.e. their crossings in order go over, under, over, under... By flipping a crossing in this figure, a non-alternating section of the diagram is produced. As we cannot increase the crossing number by flipping crossings in a minimal diagram, we cannot reach a diagram with eight or more crossings here. This means that the knot we have reached must be alternating and the non-alternating section of the diagram can be removed by Reidemeister move two. Finally, this means that any crossing flip in this figure must change the minimal crossing number by at least two. This is not the case

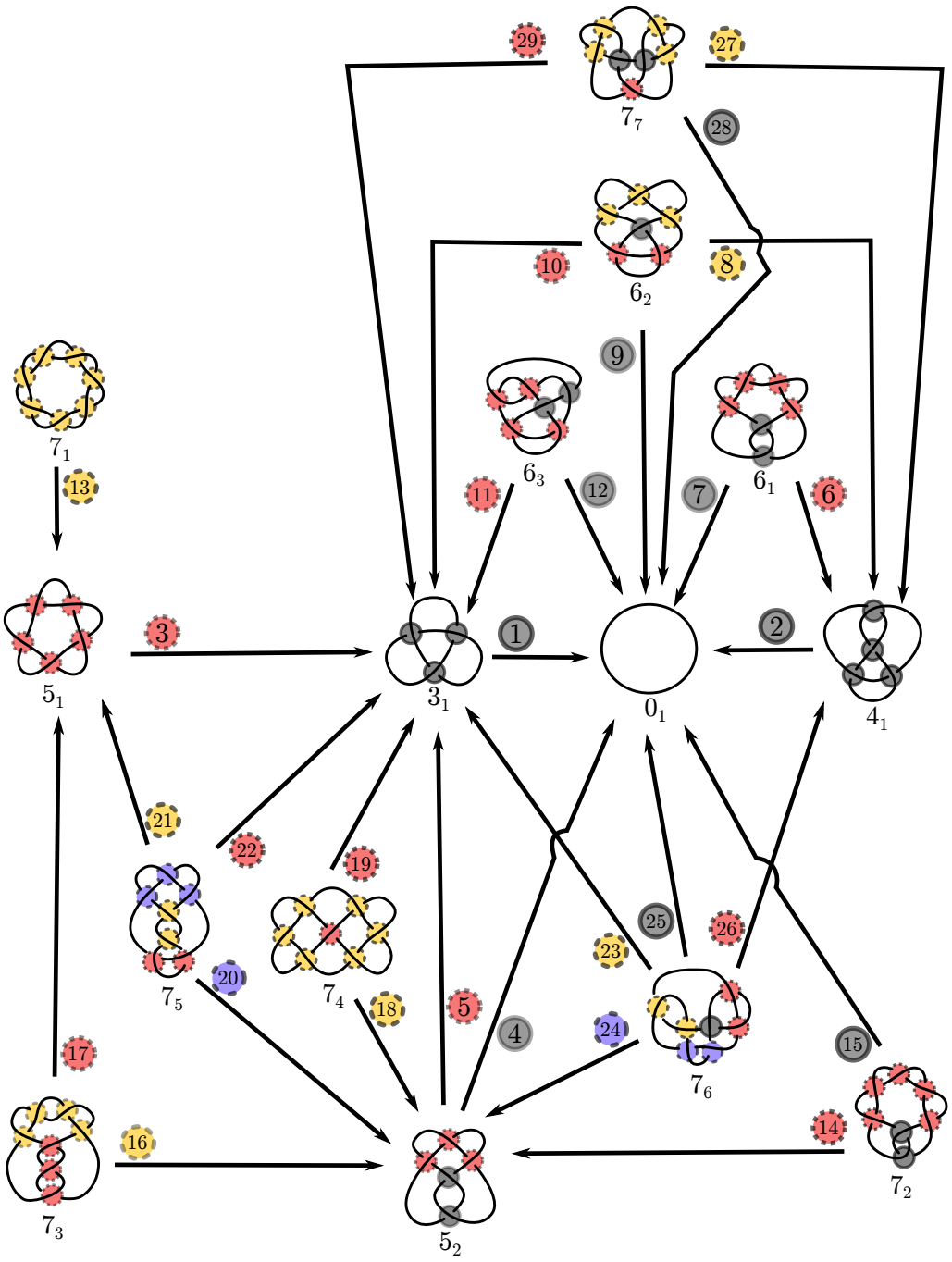


Figure 3.24: A directed graph showing the shape of knot space, up to prime knots of seven crossings. Each node of the graph represents a minimal knot diagram. A directed edge indicates that a knot can be transformed in the indicated direction to another knot by changing an over-crossing to an under-crossing or vice versa. Each directed edge has a unique label with a number and a colour. To transform a knot along a given directed edge, a crossing of the matching colour needs to be flipped. All grey crossing flips lead to the unknot. Other colours are arbitrary. If a classical crossing is replaced with a virtual crossing, the invariants of the resulting virtual knot are given in Table 3.2.

	$\lvert\Delta_g(-1,e^{2\pi i/3})\rvert$	$\lvert\Delta_g(-1,i)\rvert$	$\lvert\Delta_g(e^{2\pi i/3},i)\rvert$	$\lvert V(-1)\rvert$	knot type
1.	3	4	5	2	$v2_1$
2.	3	4	5	4	$v3_2$
3.	3	0	8	4	$v4_{100}$
4.	7	8	9	5	$v4_{43}$
5.	3	4	5	5	$v4_{94}$
6.	3	4	5	7	
7.	7	8	9	6	
8.	3	8	18	9	
9.	0	4	13	8	
10.	3	0	8	8	
11.	3	8	18	9	
12.	0	4	13	9	
13.	0	4	9	6	
14.	3	4	5	9	
15.	10	12	14	8	
16.	3	4	21	10	
17.	3	0	8	10	
18.	7	8	9	12	
19.	10	12	14	11	
20.	0	8	26	13	
21.	0	4	13	13	
22.	3	4	21	12	
23.	3	8	18	14	
24.	0	4	13	14	
25.	7	12	22	13	
26.	10	16	27	14	
27.	3	8	18	15	
28.	7	12	22	15	
29.	10	16	27	15	

Table 3.2: Table of virtual knot invariants corresponding to the directed edges of Fig 3.24. The invariants used are those described in Sec 3.2. Knot types are listed where known, and knots 8. and 11. are the only knots to share the same invariants.

when we consider non-alternating knots, where a crossing flip can change the minimal crossing number by one or even zero.

Instead, Fig 3.24 is intended as an aid to imagining the relations between knots, particularly with regards to closure analyses. Following the discussion in Sec 3.3.2, neighbouring regions in a closure analysis are most often related by a single crossing flip, and potentially by two crossing flips. For sphere closure, the neighbouring knot type regions can only be those connected by at most two directed edges in Fig 3.24. In virtual closure, a classical region can be neighboured by the virtual knots lying along those directed edges, although these virtual regions may branch off beyond Fig 3.24.

3.3.4 Virtual slipknotting

Slipknots, as described in Sec 2.2.2, are a feature of open curves characterised by the knot type under a closure analysis becoming more complicated as the curve is progressively shortened from either end, before becoming unknotted and disappearing. The work done by Millett et al. [118] in this area used a sphere closure analysis, but one could use a virtual closure analysis. This would differ from the sphere closure analysis in a few ways. We would expect that any given region of knotting during the shortening process could be reclassified as virtually knotted or weakly knotted, just as before. More interestingly, as a curve is transitioning from one knot type to another, between classically knotted regions there has to be a virtually knotted region for the same reasons as discussed earlier in this section. We would also expect that previously uninteresting unknotted regions might show virtual knotting which never become classical knotting under sphere closure.

This work has been done using *knotoids* [124, 125] and shows essentially the same features as a virtual knot analysis would show. Here, projections which we would recognise as classical knots are called *knot-type* knotoids and projections which would give virtual knots are called *proper* knotoids. In their analysis, they call regions of proper knotoids surrounded by trivial knotoids (i.e. unknots) with no knot-type knotoid region within, *pre-knots*. We will discuss a little more about knotoids in Chapter 6.

This section is mainly included for completeness regarding topological features of open curves. We will not return to slipknotting in the later chapters.

4 Virtual knots in proteins

Now that we have covered the mathematical details of detecting and classifying knots in open curves we are ready to look at an actual physical system. There is a great deal of interest in the knotted structure of proteins and much of the inspiration for this work came from studies into this [50]. As outlined in Chapter 1, the structure of proteins is intimately linked to their function and the presence of knots is a remarkable feature due to the difficulty of their formation. Given this, any methods for gaining additional structural information or insight into these proteins is valuable and it is natural that we investigate the knotting of proteins using the tools developed in the previous chapters. We begin with some addtional background on proteins covering how their structure is determined experimentally and how these details are shared in the Protein Data Bank (PDB) [126], including implications for the reliability and representativeness of the data. We then cover how we went about our knotting survey of proteins, including the cues we take from the knotted proteins database KnotProt [23], the leading contemporary knotting survey. After this, we present the results of our survey of knotting in proteins, as found in the PDB. Here we contrast sphere closure and virtual closure results, and discuss the types of knotting seen. We also look at the trend of virtual and weak knotting in different protein families, and investigate the geometric qualities of knotted proteins. Many of these results were originally published in [113].

4.1 Additional proteins background

In Chapter 1 we provided general information about protein structure and a history of the search for knots in proteins. Here we give some more detail about how these protein structures were determined and some of the limitations of

the experimental data. We also cover the PDB and how the proteins structures available there may differ from the full protein universe.

4.1.1 Experimental structure determination

While it is possible with protein sequencing techniques to determine the order of amino acids in the primary structure of proteins, this alone is not enough to tell us the final three-dimensional structure [127]. There are a number of approaches employed to probe this structure experimentally. The first protein structures to be determined were those of haemoglobin and myoglobin using *X-ray crystallography* in the 1950s [128], and this remains a popular approach. X-ray crystallography involves first making a crystal of the protein in question, and then performing an X-ray diffraction experiment, where an X-ray beam is aimed at the crystal and diffracted to form a pattern of spots. The arrangement of the spots, as well as the phase of the X-ray at each spot, is the Fourier transform of the positions of the atoms in the protein and so the structure can then be reverse engineered from this data by the inverse Fourier transform.

The effectiveness of this approach is highly dependent on how well the protein being studied crystallises, with the best data coming from crystals where each protein is exactly aligned and ordered [129]. The larger the crystal, the more sites are available for the X-rays to diffract from and so the the faster the diffraction pattern can be formed with adequate sharpness. Prolonged exposure to X-rays begins to break bonds and alter the structure of the proteins being examined and so reducing exposure time is crucial. More flexible proteins are harder to crystallize, with the flexible parts often being smeared in the diffraction pattern. Additionally, a crystalline environment is far from the conditions in which the proteins would naturally be found performing their function and so even well resolved structures must be taken with a degree of scepticism. This is a particular challenge in membrane proteins [130, 131].

X-ray free-electron lasers (XFEL) are a recent innovation on the X-ray crystallography technique. They involve using very short, high intensity pulses of X-rays to analyse the structure of protein microcrystals [132, 133]. By passing a stream of microcrystals through the pulsing XFEL to create many many diffraction patterns and later aligning and combining these patterns, the full structure of the protein may be determined. One of the advantages of this approach is that the crystals need only be very small, as long as sufficiently many are analysed, avoiding the great challenge of growing large crystals [134]. XFEL also

allows dynamic processes to be observed. For example, a light pulse can be applied to the stream moments before the XFEL pulse analyses it. By varying the timing of the light pulse before the XFEL pulse, the structural changes in response to the illumination over time can be determined [135].

NMR spectroscopy is another technique used to probe protein structure. A key difference between this and X-ray techniques is that the proteins are prepared in a pure and concentrated solution, rather than a crystal [136, 137]. This allows the protein to be studied in a more realistic environment, and has fewer demands on the rigidity of the protein, making this an effective approach to look at flexible proteins. After a purified sample of the protein is obtained, it is placed in a magnetic field and a radio frequency electromagnetic wave is applied. The magnetic field causes an energy difference in the nuclei with spin depending on if they align or antialign with the field. The lower energy aligned spins can transition to the higher energy antialigned spins by absorbing radiation of a specific frequency and thus energy. This resonance can be detected, and by sweeping through frequencies or varying the magnetic field strength, a resonance spectrum can be built up. As the resonant peak of any given nucleus depends on the molecular neighbourhood around it, with surrounding atoms creating their own local fields which the nucleus feels, the resonance spectrum gives spatial information about the protein. By combining information from NMR spectroscopy and the amino acid sequence, a model of the protein can be obtained.

A typical model will be made of many possible structures which are all consistent with the experimental data. Sections of these structures which are preserved are likely to be rigid regions, whereas flexible regions may differ in each example. This approach unfortunately struggles with large proteins, as there can be overlapping peaks in the resonance spectrum which cannot be resolved [138].

Three-dimensional electron microscopy (3DEM) is a set of techniques which all use electron microscopy in some capacity to determine structure from two-dimensional images. Within this wider umbrella, *cryo-EM* has become the most well known and used technique [139]. This involves creating a purified solution of protein, making a thin film of this solution and freezing it. The molecules are then held static in a layer of non-crystalline ice for imaging. While images are two-dimensional, the orientation of the molecules is varied and so a full three-dimensional picture can be built up. In contrast to NMR spectroscopy,

cryo-EM has been much more successful at imaging large proteins and more complex macromolecular assemblies while retaining the advantage that the proteins are in a more natural environment.

4.1.2 The Protein Data Bank

The Protein Data Bank (PDB) is an online repository of experimentally determined protein, DNA and RNA structures [126]. Each entry regards a single biological complex and contains details of the experimental technique or techniques used to resolve the structure, the publication reference for the structure and most importantly for this thesis, atomic coordinates, elements and bonds. A unique four character PDB ID is used to label each entry, with individual chains in a complex receiving a further label, typically a single character, and rarely two. In cases where an updated structure is available, the old structure is made redundant and removed from the active PDB.

For an aspiring bioinformaticist looking to use PDB data, there are a few things to be aware of. Not all PDB entries are created equally. The resolution of each entry can vary a great deal, meaning there can be large uncertainties in atomic positions. In some cases, it is not possible to determine the location of atoms that are known to form the protein from the primary structure. These missing residues are marked in the PDB files. Also, there are many very similar structures in the PDB. For example, the same protein can appear in a complex with various different compounds bound to different sites, or with only minor changes to its amino acid sequence. Additionally, the experimental challenges faced when attempting to determine protein structures have given the PDB a bias towards shorter molecules. If one is looking at statistics across the PDB, it is crucial to remember that the PDB is not a representative sample of protein space and depending on the experimental methods used, the structures may be quite distorted from their natural forms. Still, it is the best information resource available for this purpose and there is much interest in analysing the trends and patterns within.

4.2 Surveying the PDB

Before covering the results of our knotting survey of the PDB, we must cover some methodological details. In particular, how do we choose the chains we analyse, how do we parse the atomic coordinates of the protein backbones and

how do we perform the knot analysis. Also included is a brief exploration of the geometrical characteristics of the set of proteins we study, such as the chain length and frequency and size of chain breaks.

4.2.1 Selection of PDB entries to analyse

In order to make comparison of our results with existing work as simple and valid as possible, and by taking the lead of more experienced biologists, we follow the conventions of KnotProt [23, 39] in choosing which PDB entries to analyse. All protein results are for the PDB as it was in September 2016. The PDB contains structures for molecules other than proteins, such as DNA and RNA, and so these are of course omitted. The goal of KnotProt then is to build the largest set of unique protein structures to analyse. Protein structures from all experimental methods are considered, not just those from X-ray measurements. In the case of *homomultimeric* proteins, that is a protein which consists of multiple identical chains, only a single chain is analysed. Entries that contain chain breaks, where the atomic positions could not be resolved but the presence of an amino acid is known from sequence data, are still included but labelled as broken.

Gathering the exact PDB entries analysed by KnotProt is not completely straightforward, although all the information is available on the KnotProt database website if one knows where to look. In the 'Database statistics' page, accessed from the 'Read more' menu, there are two downloadable files at the bottom: 'knotted.txt' and 'unknotted.txt'. These files contain a list of the PDB IDs and chain IDs analysed by KnotProt, the complete list being obtained from the union of these two files. This list, as of September 2016, contained 159,518 unique chains, from the complete set of 329,296 chains derived from 121,532 separate protein complexes. The 'knotted.txt' files contains both knotted and slipknotted chains. The information to distinguish these is available on the 'Browse database' page. Selecting 'view raw data' gives a plain text list of all knotted and slipknotted chains, together with an indication of whether they are knotted or slipknotted, and which knot types appear in their slipknotting fingerprint. This can be parsed to separate knots and slipknots.

4.2.2 Parsing PDB files

The complete, current PDB can be downloaded readily from the PDB website in a variety of formats and file structures. We use the .pdb file format, which provides information on which atoms lie at which coordinates, which amino acid in the sequence the atoms belong to, other molecules in the structure, and other information related to the experiment and the particular protein in question. Various parsing software is available to read these files, and we use ProDy [140]. ProDy is a Python package capable of various advanced protein analysis functions, but we only use it to parse the atomic coordinates of the alpha carbon atoms, in order, of each chain. Ordinarily, the symbol of each chemical element is given in the .pdb file i.e. H for hydrogen, C for carbon etc. However, in the case of the alpha carbons, they are labelled Ca, which is also the chemical symbol for calcium. Occasionally, for proteins which are in complexes with molecules containing calcium, ProDy erroneously returns the coordinates of the calcium atoms with the alpha carbons. We have modified the ProDy code to remove this bug and only return the alpha carbon coordinates.

We join the alpha carbons in order with straight lines to produce an open curve, which we can analyse for knotting. This is an approximation to the full NCCNCC... backbone of the proteins and is commonly used in studies of protein knotting, in particular by KnotProt. In the case of chain breaks, we still use straight lines to connect the alpha carbons for which positional information is available. This is a notable difference between our analysis and that of KnotProt, which uses a more sophisticated method to model missing chain segments where possible. While this difference certainly will affect the knot analysis, cases of large breaks of more than 20Å, alpha carbons typically being 3.8Å apart, are uncommon. Further details can be found in the next section.

There are a small number of chains which do not parse successfully and so could not be analysed. In total, there are 70 such cases, all of which are reported as unknotted by KnotProt. We reported these as unknotted also in our paper [113], although strictly we only have information on 159,448 chains. As the number is small compared to the size of the PDB, the effect on statistics of the database is proportionally small also.

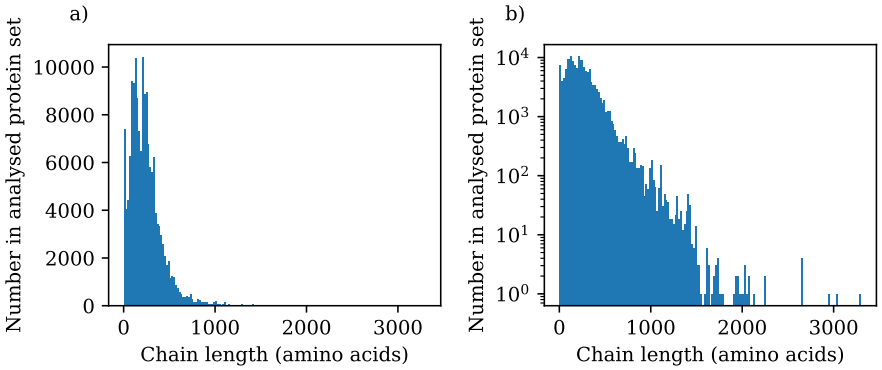


Figure 4.1: The distribution of lengths of protein chains in the chains analysed for knotting. Both plots show the same data. a) uses a linear scale and b) uses a log scale to highlight the few extremely long proteins.

4.2.3 Statistical geometrical characteristics of the PDB

As the PDB is a limited resource, constrained to those protein structures which are available to current experiments, even readers familiar with features of proteins in general may not be as familiar with the raw data we are working with. To give the reader an idea of the scale of protein backbones and how this raw data looks, we present a few plots of the key geometrical features of the proteins we analyse.

Chain length

First, Fig 4.1 gives the distribution of protein chain lengths. The vast majority of protein chains available in the PDB are under 500 amino acids long with the peak around 180 amino acids. There are a select few very long chains which are picked out in the log scale plot.

Radius of gyration

The radius of gyration of proteins is distributed similarly to length, as shown in Fig 4.2 with a strong peak at modest radius of gyration and an extended tail towards larger radius of gyration. The peak here is centred around a radius of gyration of 16Å. There were two PDB entries with a radius of gyration in excess of 800Å which have been omitted from this plot as highly unusual outliers, to better view the bulk of proteins. If we make the approximation that the mass in the protein is roughly distributed like a solid sphere, the radius of that sphere

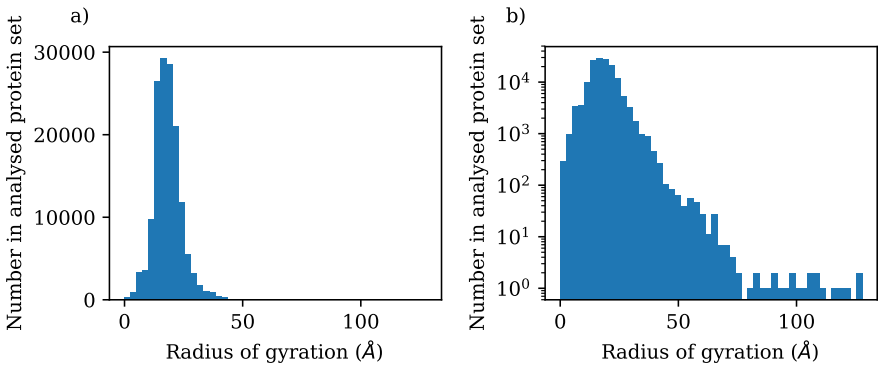


Figure 4.2: The distribution of radius of gyration of protein chains in the chains analysed for knotting. Both plots show the same data. a) uses a linear scale and b) uses a log scale to better show the tail at larger radius of gyration.

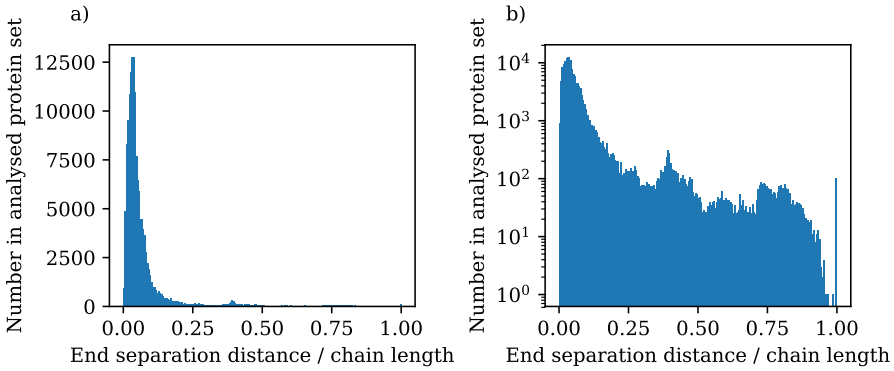


Figure 4.3: The distribution of end separations as a ratio of chain length of protein chains in the set analysed for knotting. Both plots show the same data. a) uses a linear scale and b) uses a log scale.

given a radius of gyration of 16Å would be 20Å. Given that protein lengths in the PDB are distributed around 180 amino acids and each amino acid is 3.8Å apart, we see that the proteins we analyse are highly compact structures in general.

End separation

A measure that may be important for the sort of knotting seen is the end separation of the proteins as a ratio of their total length, as given in Fig 4.3. This is in part a measure of compactness of the chains, as the closer to 1 this value, the straighter the chain. The peak of this distribution is between 0.03 and 0.04, with 90% of chains falling under 0.14.

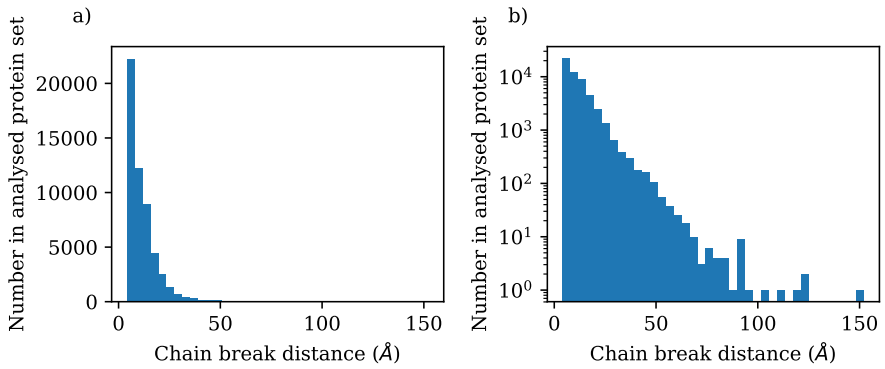


Figure 4.4: The distribution of chain break sizes of protein chains in the chains analysed for knotting. Both plots show the same data. a) uses a linear scale and b) uses a log scale to highlight the few largest breaks.

Chain break size

Finally we provide some measure of the quality of the data available in Fig 4.4, which shows the distribution of chain break sizes. That is, where one or more amino acids are missing from the atomic coordinate data, what is the size of the straight line gap bridging the alpha carbons for which we do have data. 53,694 chains analysed had gaps larger than 4Å in their backbones. Fortunately, the distribution is weighted towards the short end of the spectrum here, with 50% of chain breaks under 9.4Å and 90% under 20.1Å. There are a small number of much larger breaks, which are visible in the log scale plot.

4.2.4 Knotting analysis details

The knotting analysis that follows is performed using the methods laid out in Chapter 3. We take the alpha carbon backbone of the protein to be the open curve we analyse. We use both sphere closure and virtual closure to detect knotting and provide a comparison between the methods. 100 closures/projections are used in each case. The fractions of the knots which appear are recorded, but the directions giving each knot are not generally used. We are interested in the number of projections resulting in a given knot and not the contiguous areas of each knot type.

4.3 Virtual closure analysis of the PDB

At last we can discuss our knotting survey of the PDB. We start with our own sphere closure results and provide a comparison to the results of KnotProt before moving on to the new virtual closure results. Also included is a discussion of the geometrical properties of knotted proteins and how they compare to unknotted proteins, and information on the protein families which show particular knotting characteristics.

4.3.1 Sphere closure

KnotProt comparison

First we cover the results of our analysis of the PDB using sphere closure to detect knotting. We are, to begin with at least, comparing ourselves to KnotProt to ensure our analysis is correct. In Chapter 3 we outlined various ways of classifying the knotting seen in curves, where a strongly knotted curve gives the same non-trivial knot type in 50% or more closure directions, and a weakly knotted curve gives non-trivial knots in 50% or more closure directions, but where the most common knot does not cover a majority of closures. Hence, our classification of knotted under sphere closure is that at least 50% of closures are knots. However, this is not a convention shared by KnotProt, which recognises a chain as knotted if the most common knot type over projections is not the unknot. The most common knot is taken to represent the knotting of the curve. This is a weaker condition for knotting than our strong knotting, as the most common knot need not exceed 50% coverage, but a stronger condition than our weak knotting, as in weak knotting the unknot could be the most common knot and the curve still recognised as knotted.

We shall deal first with knotting under sphere closure according to KnotProt's knotting criteria. KnotProt reports 946 knotted chains out of the 159,518 chains analysed. We instead find 972 chains to be knotted. All but one of the chains KnotProt determines to be knotted we also determine to be knotted, leaving 27 additional proteins we find knotted which are unknotted according to KnotProt. 17 of these 27 additional detections are also considered knotted by KNOTS [37] and/or pKNOT [38]. The remaining ten chains all contain chain breaks which, as discussed, are handled differently by KnotProt, meaning the actual open curves we analyse are different. However, there is broad agreement between KnotProt and our results, and the differences can be reasonably accounted for

and are likely due to small methodological differences. Therefore we determine that our sphere closure procedure works correctly.

Of the chains KnotProt determines are knotted, 871 are 3_1, 45 are 4_1, 27 are 5_2 and 3 are 6_1. We find 894 are 3_1, 48 are 4_1, 27 are 5_2 and 3 are 6_1. Most of the additional detections are trefoil knotted, as well as the one missing knot.

Strong and weak knotting

To move beyond KnotProt at this point, we consider knotting under our definitions. As weak knotting is a weaker requirement of knotting than KnotProt's criteria, all the chains listed as knotted above will continue to be knotted here. In total, across strong and weak knotting under sphere closure, we find 975 knotted chains, 3 more than when using KnotProt's criteria. 968 of these are strongly knotted and 7 are weakly knotted. Strictly, it is only valid here to ascribe knot types to strongly knotted curves. Among these, we find 890 are 3_1, 48 are 4_1, 27 are 5_2 and 3 are 6_1. This is very similar to the knots we see using KnotProt's measure, only losing 4 trefoils to weak knotting. If we ask what the most common knots are for the 7 weakly knotted proteins, 4 are 3_1 and the remaining 3 are 0_1. These three proteins whose most common knot is the unknot would be considered unknotted under KnotProt's criteria, but as knotting is more common than unknotting in their closure spectra, they are caught by the weak knotting classification. In this way, weak knotting is more sensitive to knotting than simply asking what is the most common knot, and distinguishing between strong and weak knotting gives a degree of nuance. The results from each measure of knotting are summarised in Table 4.1.

Fraction of sphere closures which are knotted

The results presented so far give a sense of the types of knots seen, but there is more detail in the sphere closure analysis than just what the most common knot is, and whether or not it appears in a majority of projections. We can also ask how many knotted closures are there in each chain. Fig 4.5 a) shows the fraction of knotted closures on sphere closure for all proteins analysed and b) shows only the knotted proteins. Given the vast majority of proteins are unknotted, the shape of Fig 4.5 a) is no surprise. Even trace knotting is barely seen here, with almost every curve showing less than 25% knotting. Given this distribution, one may wonder if the knotted proteins are simply the tail end of this curve.

	KnotProt	Our results KnotProt's criteria	Strong knots	Weak knots
Total knotted chains	946	972	968	7
0_1	n/a	n/a	n/a	3
3_1	871	894	890	4
4_1	45	48	48	0
5_2	27	27	27	0
6_1	3	3	3	0

Table 4.1: Summary of our knotting results using sphere closure, giving the number of knotted protein chains detected under various criteria. The KnotProt column gives KnotProt's own results, and the knot type rows correspond to the number of chains whose most common knot over closures is the given knot type.

Isolating the knotted proteins shows this is not the case. Reassuringly, there is a spike in knot coverage towards the largest fractions and we see that there are very few proteins that just scrape a knotting classification. Here there are very few knotted proteins displaying a fraction of 75% or less knot coverage. There is a very apparent peak at a fraction of 0.8, and a less prominent peak at a fraction of 1. This is largely due to the small number of distinct proteins which knot. Many of the knotted protein PDB entries correspond to very closely related proteins, and many structures for these fall around the two peaks.

Focussing again on the most common knot over closures, Fig 4.5 shows the fractional coverage of the most common knot on sphere closure for all proteins analysed, c), and for only the knotted proteins, d). Again, the set of all proteins is dominated by strong unknots. Isolating just the knotted chains once more in d), first it is easy to identify the weakly knotted chains as those below a fractional coverage of 0.5. Interestingly, this graph also shows two distinct peaks. The first and largest peak is centred at a fractional coverage of 0.69, and the later, smaller peak at 0.88. Similar to the plot in b), we can attribute this to the limited set of knotted proteins available and the many entries for very similar proteins in the PDB. Given the mostly strong knotting seen in sphere closure here, we would have expected a shape like this with the most common knot accounting for a very large proportion of the overall knotting seen in each curve.

One can infer that in cases where the most common knot covers a greater fraction of closures, the knot is likely to lie deeper in the chain and is more stable to perturbations of the chain conformation. It would appear that while some knotted proteins do display a very dominant knot, many have a reasonable

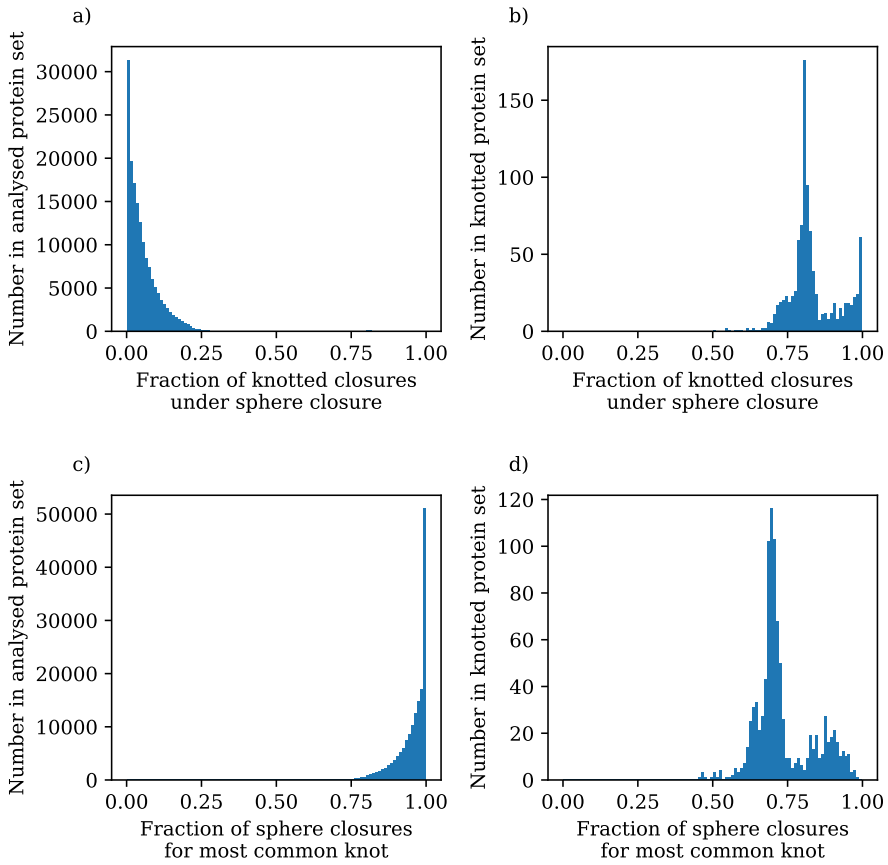


Figure 4.5: Plots showing the coverage of knots in proteins under sphere closure. a) and b) show the distribution of knotted closures, or all closures that do not return the unknot. c) and d) show the distribution of the fraction of closures which give the most common knot for each chain. a) and c) show this for all proteins analysed, which are still overwhelmingly unknotted. b) and d) show only those chains that we find to be knotted.

representation of other knots in their spectrum. This would suggest that for a number of proteins, it may not be very difficult to force a change of knot type, which given the simplicity of knots seen and the fact they all have unknotting number one, is likely to be to the unknot.

4.3.2 Virtual closure

Strong and weak knotting

We now move on to deal with our virtual closure results. An important point to remember at the start here is that all chains detected as knotted under sphere

closure are also detected as knotted under virtual closure, although the classification of this knotting may change. In total, we find 1,258 chains to be knotted under virtual closure i.e. the unknot appears in 50% or fewer virtually closed projections. This is 283 more chains than under sphere closure, indicating that virtual closure is a more sensitive knot detection method.

In comparison to sphere closure, we now have five different classifications for knotting instead of just strong and weak, as outlined in Chapter 3. As a reminder, strong knotting is now split into strong classical and strong virtual knotting, depending on whether the knot which dominates is classical or virtual. Weak knotting is split into three categories: weak classical, where classical knots hold a majority; weak virtual, where virtuals hold a majority; and weak total, where neither classical nor virtual knots hold a majority, but all knots together have at least 50% coverage.

Most of the knotted proteins found fall into strong classical knotting, with 727 cases. All of these were also strongly knotted under sphere closure and include 660 3_1, 46 4_1, 19 5_2 and 2 6_1 knots. As these counts are all lower than under sphere closure, examples of all knot types previously seen must be reclassified under virtual closure. An example of a strong knotted chain is given in Fig 4.6 a).

Strong virtual knotting is much less common than strong classical knotting as we find only 41 examples, 11 of which were detected under sphere closure. In all but one of these, the virtual trefoil $v2_1$ dominates, as in Fig 4.6 b). The remaining chain contains a $v4_{43}$ knot, which can be thought of as lying in between the unknot and 5_2.

The remaining chains are all weakly knotted. 343 are weakly virtually knotted, see Fig 4.6 c), of which 102 were knotted under sphere closure, and 145 are weakly totally knotted, with all but 15 of those also being knotted under sphere closure as in Fig 4.6 d). Only 2 chains are weakly classically knotted, both of which were knotted under sphere closure, one of which is shown in Fig 4.6 e). Most new detections then fall into the weak virtual knotting category, and most reclassifications are weak total knotted. These results are summarised in Fig 4.7.

Table 4.2 shows the most common knot over virtual closures for each category of knotting. Strong knotting we have already covered as these are the chains with a well defined knot type. While the most common knot is less meaningful for weakly knotted chains, it is nonetheless interesting to see which knot type was closest to representing the curve.

There are several notable features revealed in this table, starting with the

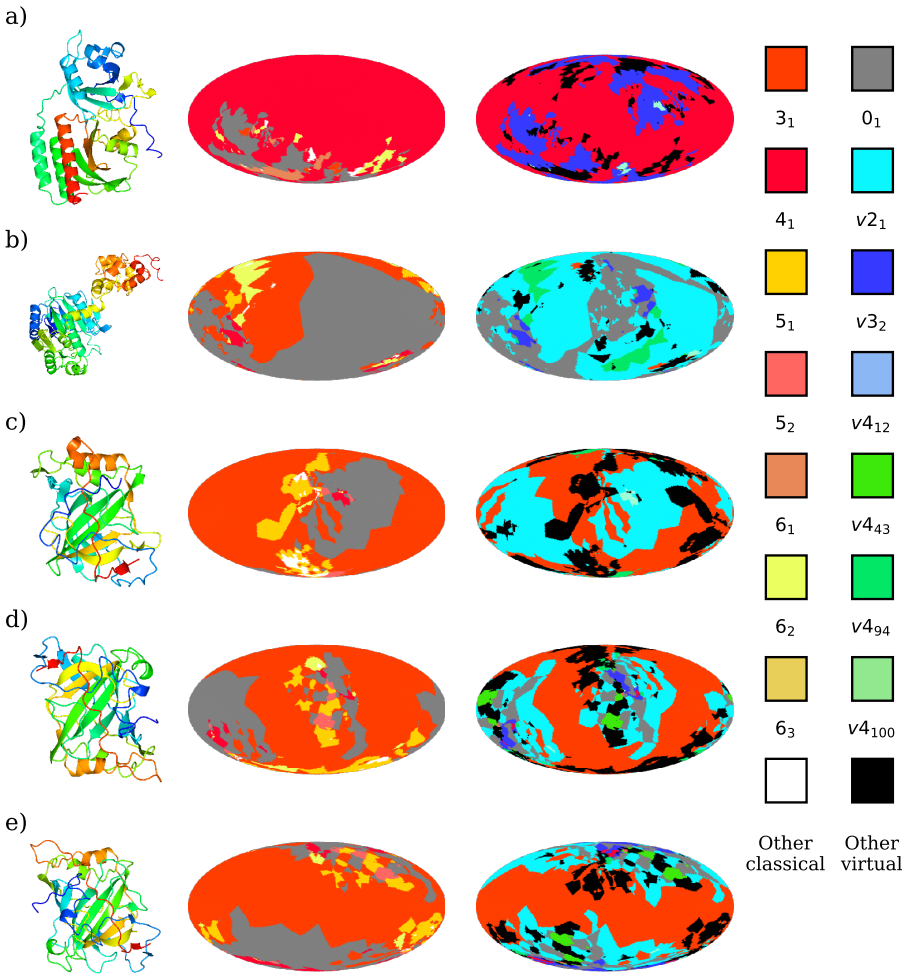


Figure 4.6: The ribbon diagram, map of the knot globe on sphere closure and map of the virtual closure globe for a selection of knotted proteins. a) PDB ID 4E04, chain A [141]. Strongly 4_1 knotted under both sphere and virtual closure. b) PDB ID 3WKU, chain B [142]. Unknotted under sphere closure, but strongly $v2_1$ knotted under virtual closure. c) PDB ID 4XIX, chain A [121]. Strongly 3_1 knotted under sphere closure, weakly virtual knotted under virtual closure. d) PDB ID 3KIG, chain A [143]. Strongly 3_1 knotted under sphere closure, weakly total knotted under virtual closure. e) PDB ID 1CZM, chain A [144]. Strongly 3_1 knotted under sphere closure, weakly classical knotted under virtual closure. The full globes and graphs of connected areas for these proteins are given in Appendix A.

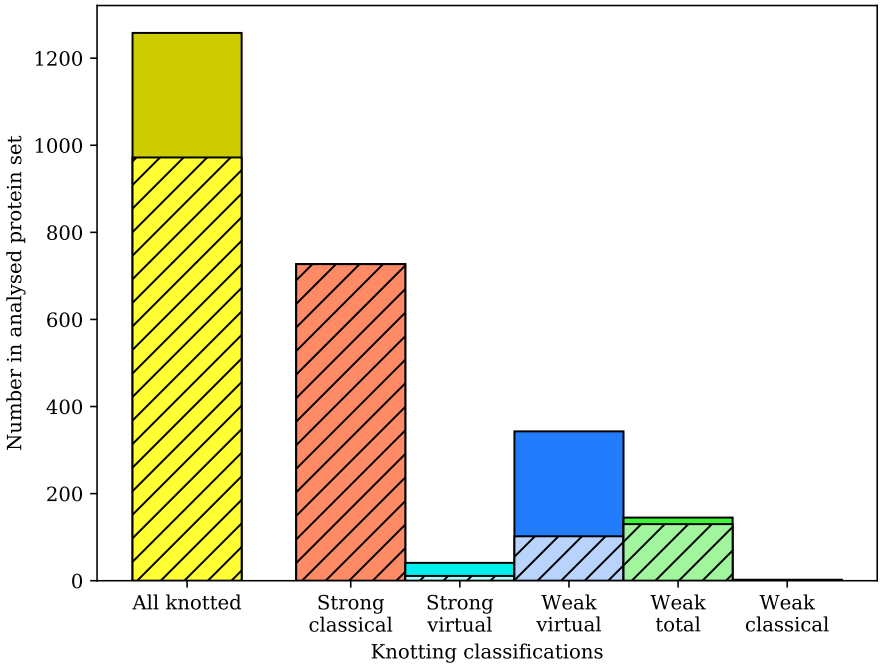


Figure 4.7: The number of protein chains falling into each knotting category using virtual closure. The hashed bars are chains which are detected as knotted also under sphere closure.

	Strong classical	Strong virtual	Weak virtual	Weak total	Weak classical
Total knotted chains	727	41	343	145	2
0_1	n/a	n/a	215	15	0
3_1	660	n/a	82	127	2
4_1	46	n/a	2	0	0
5_2	19	n/a	3	3	0
6_1	2	n/a	1	0	0
$v2_1$	n/a	40	35	0	0
$v3_2$	n/a	0	4	0	0
$v4_{43}$	n/a	1	1	0	0

Table 4.2: Summary of our knotting results using virtual closure, giving the number of knotted protein chains detected under various criteria. The knot type rows correspond to the number of chains whose most common knot over closures is the given knot type.

unknot being the most common knot in the majority of weakly virtually knotted chains. We might have expected this given the large number of weakly virtually knotted chains which were classed as unknotted under sphere closure. Thinking of the knot globe, as we move to virtual closure, areas of unknotting are eaten away by areas of virtual knotting. In the case of the newly detected, weak virtual knots this was just enough to tip them into a knotted classification. There must also have been multiple different virtual knot types in each of these for them to have avoided strong virtual classification.

A similar change takes place for those curves previously classed as trefoil knotted, losing regions of trefoil knotting to virtual knotting. While the trefoil remains the single most common knot, virtuals of different types have collectively taken over the knot spectrum.

There are surprisingly few weakly virtually knotted curves where a virtual knot is the most common knot. The majority where this is the case show a $v2_1$ plurality and are on the verge of strong virtual knotting.

Similar readings can be made into the weak totally knotted chains. Most of these were strong trefoil knots under sphere closure, but some of the trefoil knotting has been deposed by virtual knots under virtual closure. Not enough so that the chains become weak virtual however. The handful of new detections still display the unknot as the most common knot, but there must have been a significant classical knotting fraction present under sphere closure also, as virtual closure cannot introduce new classical knotting regions.

The two lone weak classically knotted chains are almost trefoil knotted, having been strongly trefoil knotted under sphere closure. Virtual knots have removed the majority held by trefoil but not enough to erode the classical flavour present.

One thing that is apparent from the knot maps shown in Fig 4.6 is that our distinction between weak classical, weak virtual and weak total knotting does not translate to very different knot profiles. The distinction between strong and weak knotting is much more dramatic and so we find it meaningful to group all strong knotting together, and all weak knotting together in much of the analysis which follows.

Fraction of virtual closures which are knotted

We can also look at the distribution of knotted fraction sizes as we did for sphere closure, shown in Fig 4.8. This can give us an idea as to the stability of

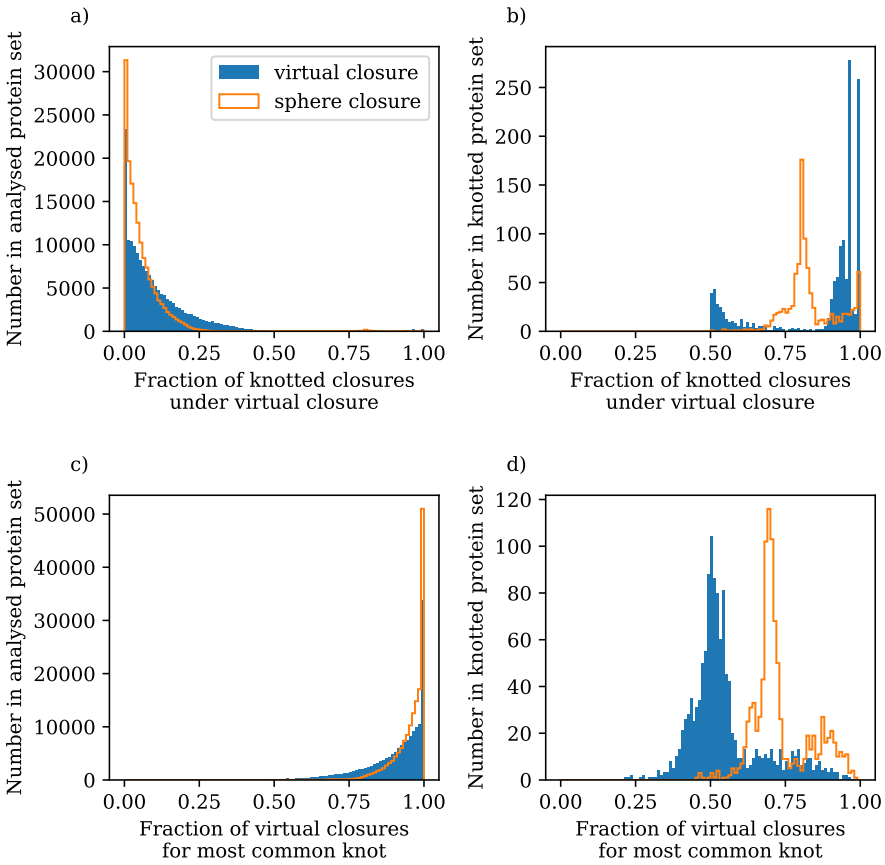


Figure 4.8: Plots showing the coverage of knots in proteins under virtual closure in blue, and sphere closure in orange outline. a) and b) show the distribution of knotted closures, or all closures that do not return the unknot. c) and d) show the distribution of the fraction of closures which give the most common knot for each chain. a) and c) show this for all proteins analysed, which are still overwhelmingly unknotted. b) and d) show only those chains that we find to be knotted.

knottedness to perturbations and show how sphere and virtual closure differ. Looking at the set of all proteins where the unknot dominates we again see that the distribution is highly weighted towards high unknot coverage, but the tail is broader than what was seen under sphere closure. More trace knotting is detected using virtual closure. This is true when considering all knots, a), and just the most common knot, c).

Examining the fraction of knotted closures for the knotted chains, in Fig 4.8 b), we see that there is a large concentration of proteins with a knot coverage of over 90%. This is considerably more knot coverage than was seen for most proteins under sphere closure. We also see a feature here not evident under sphere clo-

sure, and that is the presence of knotted proteins at the end of the tail seen in a). Virtual closure typically amplifies the amount of knotting detected, compared to sphere closure, pushing all chains towards more knot coverage. This new knotted tail area almost seems like an accident of the broader distribution and these proteins with a knotted fraction less than 75% are clearly distinct from those with a higher knotted fraction.

Looking to the coverage of the most common knot, d), we see a fairly broad peak centred at a fraction of 0.51. This is significantly lower than under sphere closure, with many chains hovering just above a weak knotting classification. In some ways, this highlights the arbitrary nature of a 50% cut off for strong knotting. Is a 51% most common knot coverage very different from a 49% coverage? In terms of conformations of curves, not likely. Neither of those scenarios show a particularly strong knotting character, where one knot can be said to unambiguously represent the knotting of the curve. The second, higher fraction, peak seen in sphere closure appears now to be smeared out to lower coverages. For many proteins under virtual closure it seems that while the entangledness is not under question, an exact knot type is less forthcoming. This makes the presence of the proteins with a large representation of their most common knot more remarkable. These are chains which are already rare for being knotted, and are even rarer among their peers for having a well defined knot type.

In general then, we see that virtual closure increases detections of knotting, and increases the fraction of closures determined as knotted, but simultaneously decreases the certainty of a representative knot type as it introduces many new possible knot types.

4.3.3 Families of knotted proteins

A key biological question about knotted proteins is which proteins knot? We have alluded to already that the knot statistics presented are influenced heavily by a select few proteins for which there are many PDB entries. Here we will cover this in more detail.

Doing this from the PDB data itself is a little difficult however. There is limited information in the files about the broader biological context of the proteins and it is not always presented consistently. For example, some proteins are referred to as ‘carbonic anhydrase 2’ while others are referred to as ‘carbonic

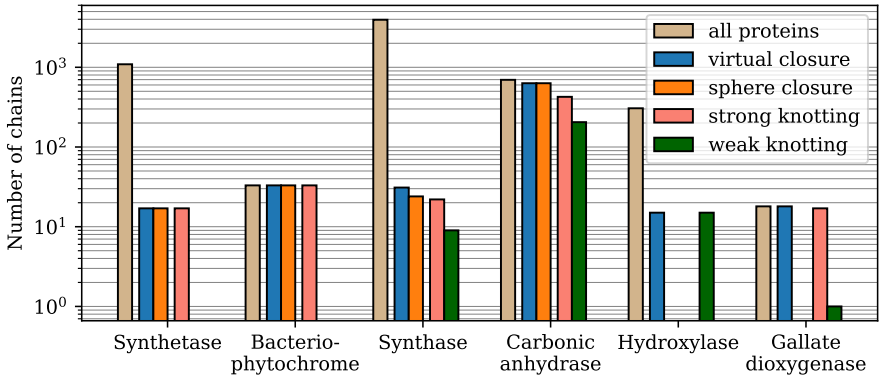


Figure 4.9: A selection of protein families which show knotting. Note the log scale.

anhydrase II'. We will be looking at the 'name' field in the PDB files and trying to catch the examples where names are inconsistently used, mostly by hand.

Of those chains which we detect as knotted under virtual closure, there are 292 different protein name strings. The two most common are 'carbonic anhydrase 2' and 'carbonic anhydrase II' with a total of 534 examples. The frequency of each name in the knotted proteins drops off very sharply after this with the next most common name, bacteriophytochrome, occurring only 29 times.

In Fig 4.9 we have chosen a select few particularly interesting names from a knotting perspective. The values here are obtained by searching for the given string in the name field. For example, 'carbonic anhydrase' combines 'carbonic anhydrase 2', 'II', '1', '13' and so on (these are the four most common knotted carbonic anhydrases in order). We give the total number of such proteins in the analysed protein set, those knotted under virtual and under sphere closure, and the distribution between strong and weak knotting under virtual closure.

The knotting seen in *synthetases* is not particularly common compared to the overall number of synthetase PDB entries. However, the knotting seen is strong and there are no new detections under virtual closure, so those synthetases which do knot are most likely deeply knotted. *Bacteriophytochrome* proteins show a similar knotting profile, but with every example available being knotted. These entries do come from different experiments and different groups, but of course each entry is highly related, and there is only a small number of samples available.

Synthases are the most common family of proteins to show knotting, and

like synthetases are mostly unknotted with only a few knotted examples. Most of these knots are caught under sphere closure but a number of weakly knotted chains are identified under virtual closure.

As previously mentioned, *carbonic anhydrases* dominate the knotting statistics and the sheer number can be seen here. Almost every carbonic anhydrase example is knotted and detected as such under both virtual and sphere closure. The knots seen are mostly strong, although there is a sizeable minority which are weakly knotted. This is not unexpected given the initial dismissal of carbonic anhydrases as only shallowly knotted. It is significant however that none were shallow enough to be missed by sphere closure.

Hydroxylases are particularly interesting from a virtual closure perspective. Under previous surveys, no knotting was detected in this family but here we find a small fraction to be weakly virtually knotted. All of the knotted structures are from the same paper [145], each in a complex with iron and another compound. It is likely then that this knotting is not a common state for hydroxylases and is brought on by environmental factors related to the experiments conducted.

Finally, every example of *gallate dioxygenase* in the PDB shows knotting only under virtual closure. Most of this knotting is strong and it is interesting that a strongly virtually knotted chain would be undetected as knotted under sphere closure. Similar to already discussed proteins, these structures all come from the same paper [142] and so more structures are needed to determine whether knotting is a generic feature of these proteins or specific to these experiments.

Many of the other knotted chains found are the sole representatives of their families, many such families not having many structures available to analyse. It is difficult to say without more specific biological information how significant the knotting there may be. Highlighted here really is how unrepresentative a data set the PDB is, as remarkable as it is in other ways.

4.3.4 Statistical geometric characteristics of knotted chains

We now turn our attentions to the broader geometrical aspects of the knotted proteins, just as we previously looked at these aspects for the entire PDB. Are there overall geometrical differences between knotted and unknotted proteins, between proteins knotted under sphere closure and virtual closure, and between strongly and weakly knotted proteins?

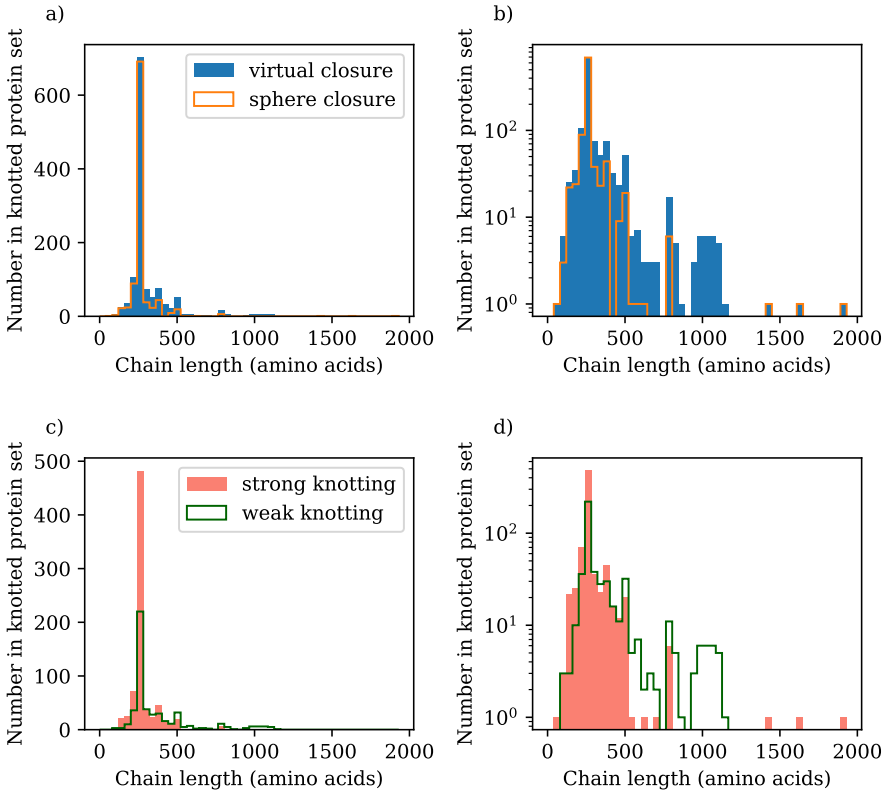


Figure 4.10: The distribution of lengths of knotted protein chains. Plots a) and b) compare chains knotted under virtual closure to those under sphere closure. Both plots show the same data. a) uses a linear scale and b) uses a log scale to highlight the few extremely long proteins. Plots c) and d) show only virtually knotted proteins and compare strongly knotted (strong classical and strong virtual) to weakly knotted (weak classical, weak virtual and weak total) chains, c) with a linear scale and d) with a log scale.

Chain length

The distribution of knotted chain lengths is given in Fig 4.10, with a) and c) using a linear scale and b) and d) a log scale. Immediately apparent is the huge fraction of knotted chains with length around 265 amino acids. Almost all of these turn out to be carbonic anhydrases. In this way, the distribution of knotted chain lengths is dominated by the number of PDB entries available for carbonic anhydrases. The broader distribution isn't too dissimilar to that of all proteins in the PDB, although there is a shift here towards shorter proteins.

Comparing those proteins knotted under sphere closure and under virtual closure in a) and b) we see broad agreement between the sets. There are a number of new detections below 500 amino acids. More interestingly there are

a number of new detections around 900 to 1200 amino acids where previously no knotting had been recognised. Given the relatively small number of such curves, it is almost certainly due to a quirk of the PDB structures available, with these chains being from closely related proteins.

Looking at the contrast between strong and weakly knotted chains in c) and d) we see similar things. Broadly, strong and weakly knotted chains have a similar length distribution with a tendency for weakly knotted chains to be longer. We see the same excess of weakly knotted chains between 900 and 1200 amino acids as under virtual closure, as these new detections must largely have been weakly knotted. We draw the same conclusions that this is likely a feature of the PDB and cannot comment about open curves in general from this data.

Radius of gyration

Looking to the radius of gyration of knotted proteins in Fig 4.11, we see again a peak around 16Å. Compared to the complete PDB distribution, there is, like length, a shift towards smaller radii of gyration with notably fewer large outliers. It is not unexpected that knotted proteins are more likely to be compact than generic proteins, as relatively short tangled structures demand a certain compactness.

We actually see relatively little difference between those proteins knotted under sphere closure and those under virtual closure. The majority of new detections lie above the peak radius of gyration but are fairly evenly distributed there, mostly serving to smooth out the rougher tail seen under sphere closure. Looking at strong and weak knotting in c) and d) we see essentially the same features as picked out by the sphere closure to virtual closure change, with no obvious difference in radius of gyration between weakly and strongly knotted curves.

End separation

The distribution of end separation as a fraction of chain length, given in Fig 4.12, shows a number of differences from the complete PDB. First, the range of the x-axis here is much shorter than for Fig 4.3. The very large end separation to chain length fractions in the whole PDB mainly arise from very short chain fragments, and these do not have enough length in order to form knots. The peak of the knotted proteins is close to that of the whole PDB, but it is much

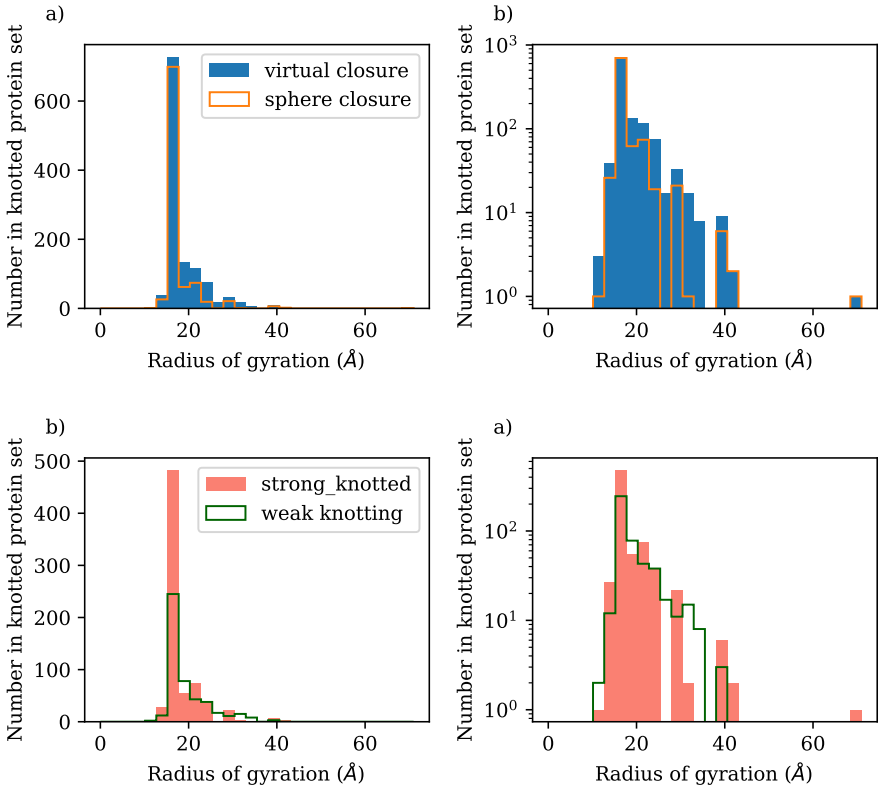


Figure 4.11: The distribution of radius of gyration of knotted protein chains. a) and b) contrast knots under virtual and knots under sphere closure. Both plots show the same data. a) uses a linear scale and b) uses a log scale. c) and d) contrast strong and weak knots, c) with a linear and d) with a log scale.

sharper for knotted proteins. Again, this is due to the many carbonic anhydrases which are knotted.

Comparing the knots under sphere and virtual closure in a) and b) we find that all the new detections are grouped at the smallest fractional end separations, broadening slightly the peak seen under sphere closure. These chains are more likely to have one end partially embedded in the bulk of the curve, perhaps not fully threading the loop required to form a knot. In this way, they avoid detection under sphere closure, but present virtually knotted projections which we now capture.

Interestingly, when we look at the end separation of strong and weak knots in c) and d), we see that the outline of the distributions is fairly similar, accounting for the large carbonic anhydrase peak in strong knotted chains, but with the weakly knotted distribution sitting closer to a smaller end separation to chain

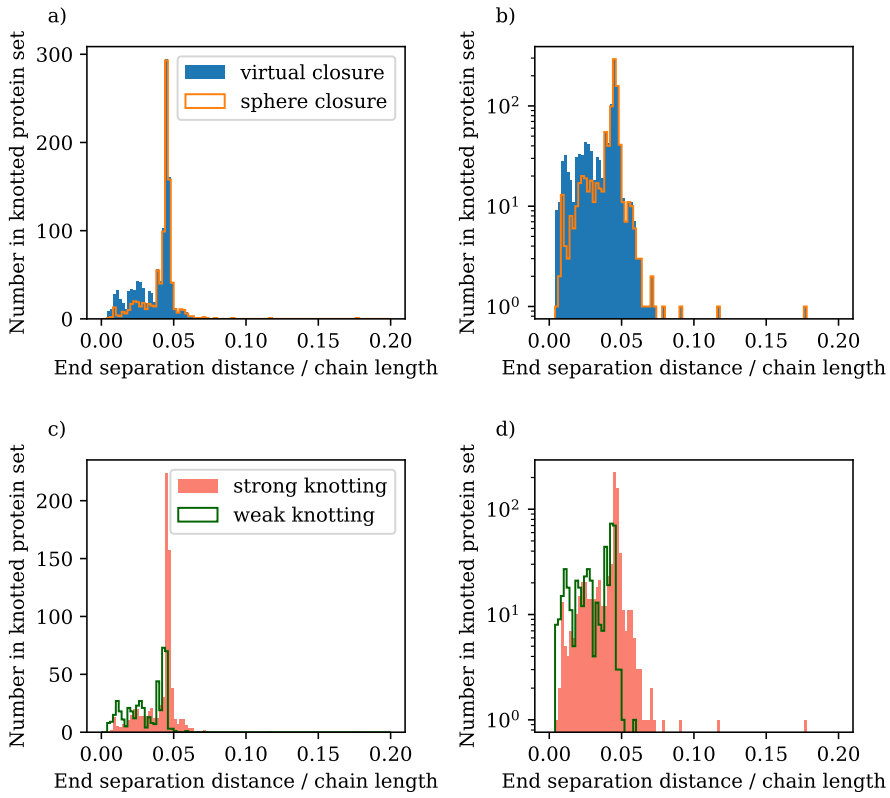


Figure 4.12: The distribution of end separations as a ratio of chain length of knotted protein chains. Plots a) and b) show knotting under virtual and sphere closure, a) with a linear scale and b) with a log scale. Plots c) and d) show strong and weak knotting, c) with a linear and d) with a log scale.

length fraction. Following similar reasoning to the contrast between virtual and sphere closure, this could be arising from the partial burying of one or both ends, bringing the end separation down and encouraging an ambiguity in knot type over closures, hence giving weak knots. Conversely, a large end separation makes the ends more likely to lie further from the bulk, resulting in most closures giving the same knot type and hence strong knotting.

We include here also the analysis of knot probability against end separation completed for our paper [113]. At the time we were trying to highlight virtual knotting as opposed to weak knotting, and the data is no longer available to replot this with our new focus. Nevertheless, we expect that the curve for virtual and for weak knotting would be fairly similar. Fig 4.13 shows how the probability of classical and virtual knotting varies with end separation for proteins, but also two different random walk models. While the specifics of these graphs vary,

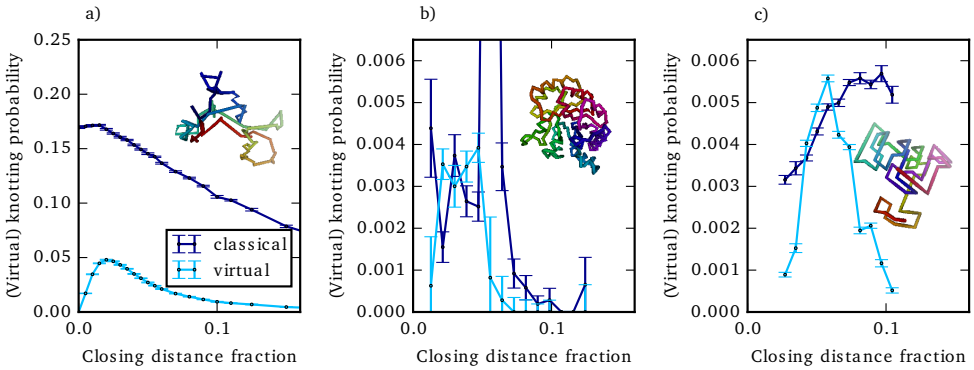


Figure 4.13: The probability of knotting and virtual knotting with end separation as a ratio of chain length for a variety of open curves. Plot a) shows data for off-lattice random walks, b) for protein chains and c) for lattice walks. The details of the random walks will be discussed more in the following chapter. This figure is taken from [113].

what they all have in common is a peak of virtual knotting at roughly the same end separation as a fraction of length, between 0.02 and 0.05.

Closest end-point to centre of mass distance

Another quantity we expect to be significant for weak knotting in particular is how close one of the end-points is to the centre of mass of the chain. We plot the distribution of this distance for knotted proteins, as well as its effect on weak knotting probability in Fig 4.14. As we see from a), the end-points of proteins are unlikely to lie close to the centre of mass of their backbones as we expect [21]. One would think that having significantly buried end-points would increase the chance that knotting is weak, but this is not clear in b). We will investigate this further in the next chapter.

Chain break size

An important question to ask is are the knotted proteins merely artefacts due to chain breaks? Fig 4.15 shows the distribution of chain break sizes for the knotted proteins. The shape is not too dissimilar to that seen for the entire PDB, and most breaks are under 20Å in size. However, the distribution does seem to be a little shifted towards longer breaks. Many of the new detections using virtual closure have short breaks also. We should therefore be careful in reading too much into the exact results described in this section, as more refined future data may change some things slightly. Given that the size of breaks generally

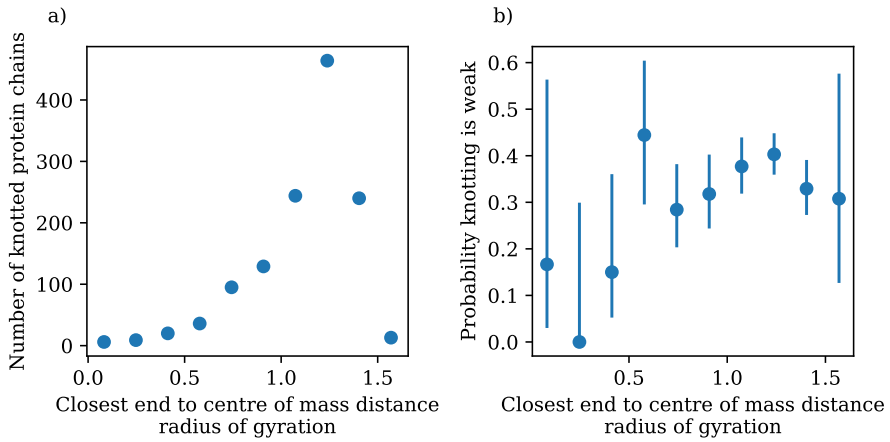


Figure 4.14: a) the distribution of the closest end to centre of mass distance as a ratio of radius of gyration for knotted protein chains. b) the probability of weak knotting with closest end to centre of mass distance as a ratio of radius of gyration.

is small however, it is unlikely that the knotted character of many proteins is completely wrong.

4.3.5 Reflections on strong and weak knotting

While the PDB may be far from a representative sample of proteins, with biases in the proteins which can be studied experimentally and typically the structures available being crystalline rather than the true *in vivo* forms, we can attempt to draw conclusions about knotting in open curves from what we see. Proteins are clearly very special curves and the proportion which fall under each category of knotting we have outlined will likely differ for other systems, one striking feature is the rarity of weak classical knotting. Under sphere closure, only 7 of 975 knotted chains were weakly classically knotted. This drops to just 2 chains of 1258 under virtual closure. While there are no other possible weak categories under sphere closure, there are 488 other weakly knotted chains under virtual closure. This suggests that there is a relatively small range of conformations which result in weak classical knotting.

Let's think about why this might be. For a curve to be weakly classically knotted it must not have a single dominant knot type over closures. It must also have a a majority of classically knotted closures, and so a representation of different classical knot types in its knotting spectrum. Considering just virtual

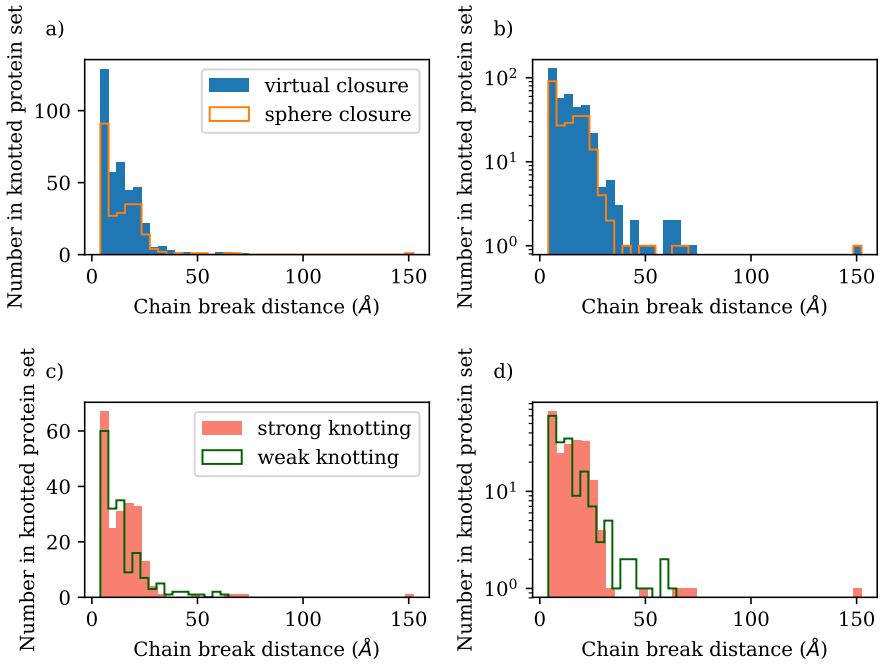


Figure 4.15: The distribution of chain break sizes of knotted protein chains. Plots a) and b) compare those chains knotted under sphere closure to those under virtual closure. a) uses a linear scale and b) uses a log scale, both show the same data. Plots c) and d) deal only with those chains knotted under virtual closure, comparing strongly knotted chains and weakly knotted chains. c) uses a linear scale and d) a log scale, both show the same data.

closure, between these different classical knot types in the globe of closure directions there very likely lie virtually knotted regions. In simply knotted curves like proteins, there may be unknotted regions also, with more virtual knotting between these and the classical regions. Given the inevitability of virtually knotted regions in weakly classically knotted curves, they must tread a fine line between possessing multiple distinct classical knot types, and not giving too much area to virtual knots, tipping them into weak total knotting, or even weak virtual knotting. It seems from the protein data that this balance is very difficult to strike.

Considering sphere closure instead, the problem of virtually knotted intermediate regions is not present, yet we still do not see much weak knotting. A few factors may play into this. The first is that the knots seen in proteins are relatively simple and so there are relatively few possible knot types that could appear in a closure analysis, given there are only seven classical knots with six or fewer crossings, not including the unknot. The second is that there is

more conformational difference between classical knots than there is between classical and virtual knots. For example, the trefoil and the figure-eight are more different from each other than the trefoil and $v2_1$ are. Hence we expect that it would be more likely for a curve to show significant fractions of both trefoil and $v2_1$ under virtual closure than trefoil and figure-eight under sphere closure.

We cannot tell from this data if the trend of weak knotting being uncommon under sphere closure will be a common feature of all open curves. We will investigate this further in the next section.

5 Confined random walks

While the results discussed in the previous chapter were a proof of concept for the virtual closure method, and do provide additional detail over existing sphere closure results, they lack context as we know that protein knots are unlike knots in other more random systems. Are the virtual knots, and in particular the weak knots seen in proteins typical for a compact open curve or atypical? To provide this context we here look at knotting in random walks, which have been studied as models for less structured polymers like DNA as well as for their own sake. In particular, we will be investigating the effects of confinement on the knotting of these walks which will encourage them to be compact[1] and in some ways more comparable to protein conformations. Also, in comparison to unconfined walks, we expect that the compactness will encourage differences between sphere and virtual closure, generating ambiguity in knot type and allowing us to better understand what each method is telling us.

We will first cover how we generate the random walks we investigate, paying particular attention to the confined walks. We confine lattice walks to a cube, and off-lattice walks to spheres, tubes and between two parallel planes, or slits. In generating the lattice walks in cubes and off-lattice walks in spheres we use pre-existing algorithms. However, we had to develop our own algorithms for the walks in tubes and slits, and we present how we accomplished this, ensuring a uniform density of vertices throughout the confining volumes. Then we present the results of our investigation into the knotting of these walks, including how sphere and virtual closure differ, the effect of the different confining geometries and random walk models, and emphasise why we think that weak knotting is an important quantity that we can measure and understand. The results are

[1] The confined walks will be more compact than unconfined walks in the parameter ranges investigated here. Very constricted tubes and slits will in fact have the opposite effect.

Figure 5.1: Examples of confined lattice walks made from segments of Hamiltonian walks. The underlying lattice has $6 \times 6 \times 6$ nodes and the number of steps in each walks is 25, 60, 94, 129, 163. This means the fractions of nodes which the walks reach are 0.12, 0.28, 0.44, 0.6, 0.76.

currently in preparation for journal submission.

5.1 Generating confined random walks

In our investigations of knotting in random walks we looked at knotting both on and off lattice. For the off-lattice case we generate walks both with and without confinement, whereas for the lattice walks we only generate confined curves. Here we describe how we create these walks.

5.1.1 Lattice walks

Our lattice walks lie on a cubic lattice of finite extent i.e. $L \times L \times L$ lattice points. The underlying lattice here is cubic as well as the shape of the confinement. From the algorithm used to generate them, they are also naturally self-avoiding, guaranteeing that we can ask about their knotting. Examples of walks of various lengths are given in Fig 5.1.

This kind of walk was used as a model of proteins by Lua and Grosberg [146] where they were shown to exhibit local geometrical similarities, but some important differences also. Looking at segments of walks and proteins of small lengths, the end separations of these were both seen to start of increasing before plateauing at longer lengths, showing that the walks and proteins both exhibit a similar compactness. The details of this trend in proteins are important however. It was seen that at very short lengths, up to 10 amino acids, the proteins were more straight than the random walk segments. This was followed by a range of lengths up to 40 amino acids where the end separation grew very slowly compared to the random walks, indicating a tendency for the protein chains to fold back on themselves.

Also considered was the degree of *interpenetration* of the proteins and random walks. Interpenetration here is a measure of how much of the rest of the chain, on average, is nearby a subchain of a given length, as the subchain being considered is taken from all parts of the protein or random walk. The confined lattice walks and the proteins both showed an increase and then plateau of interpenetration as length of subchain increased, but the plateau was at lower values of interpenetration for proteins.

From these measurements, Lua and Grosberg concluded that proteins are more 'segregated on the intermediate scale' than random compact walks. In the extreme, a highly segregated space curve would be a tight concertina or zigzag lying in the plane, or perhaps folded carefully on top of itself in another zigzag. Such a conformation, while compact, will not be knotted. For this reason, Lua and Grosberg reason that proteins are less knotted than random walks, a fact also borne out in their data.

The algorithm we use to generate these walks is taken from work by Lua, Borovinskiy and Grosberg [22], which in turn was inspired by Ramakrishnan et al. [147]. The implementation in Python we use was written by Dr. Alexander Taylor. I extended this to arbitrary cuboidal lattice dimensions but ultimately this was not used. The walks are based on *Hamiltonian walks* on the graph structure underlying the lattice, a Hamiltonian walk being a path through a graph which visits every node or vertex only once [148].

These walks were investigated by Lua et al. [22, 146] for $L \times L \times L$ cubic lattices, but the algorithm will work on any *finite, regular bipartite* graph. Finite meaning not an infinite number of graph nodes, regular meaning that each node is connected to the same number of other nodes, and bipartite meaning that each node can be given one of two colours such that no connected pair of nodes are the same colour, like a chess board.

The details of the lattice we want our walk to lie on will influence how the walk is generated. The lattice is *even* if there are an even number of lattice points, and *odd* otherwise. For an $L \times L \times L$ cubic lattice, the number of lattice points, and hence the length of a Hamiltonian walk on the lattice, is L^3 and so the lattice is odd if L is odd, and even for even L. This has implications for the end positions of the walks. Picture the lattice coloured black and white like a chess board. Any Hamiltonian walk on an even lattice will start and end on different colours, whereas the ends of a walk on an odd lattice will lie on identical colours. Further, in an odd lattice, one of the colours will always be

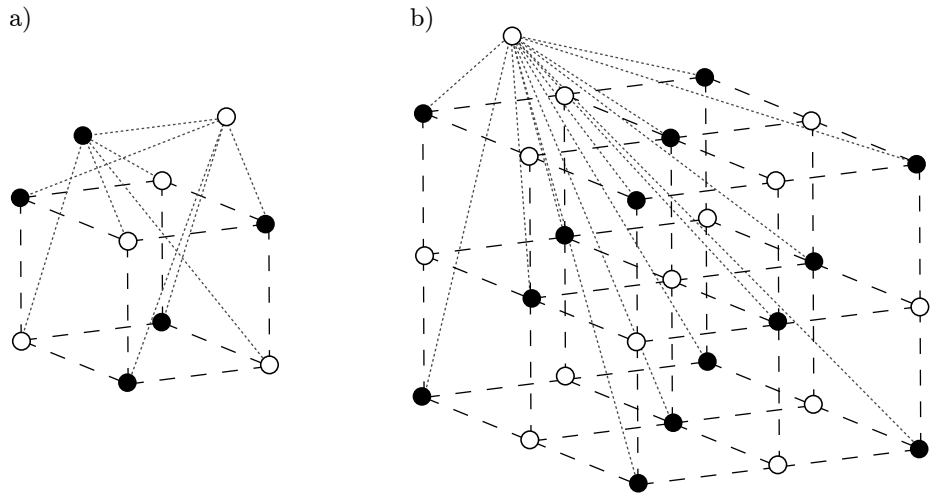


Figure 5.2: The extended lattices used to build the lattice walks. An even lattice is shown in a), and an odd in b). The coarser dashed lines show the connections between the underlying cubic lattice, and the softer dashed lines the connections to the extended lattice points.

more common than the other. We call this the *major* colour. A consequence of this is that *Hamiltonian cycles*, which are Hamiltonian walks that return to their starting point, are impossible in odd lattices as any closed loop on a bipartite graph must contain equal numbers of each colour

In order to tackle the seemingly separate problems of Hamiltonian walks on odd and even lattices, and Hamiltonian cycles in even lattices, Lua, Borovinskiy and Grosberg employ a trick used by [147]. If the lattice is even, two additional out-of-lattice points are added, one of each colour. These are connected to each other, and to all points of opposite colour in the lattice also. If the lattice is odd, one additional point of the *minor* (less frequent) colour is added out-of-lattice and connected to all the major coloured points in lattice. Fig 5.2 shows the resulting extended lattices. These extended lattices are always even and the Hamiltonian walks in lattice can be constructed by generating Hamiltonian cycles on the extended lattices and removing the out-of-lattice points.

Given the extended lattice, the algorithm proceeds by generating the initial Hamiltonian cycle. This is done essentially by randomly linking connected lattice points. If the non-extended lattice is even, the first pair connected is always the two out-of-lattice points. If the link would create a *subcycle* on the lattice or a *dead end* as in Fig 5.3, it is rejected. Once every lattice point has two links, the lattice is *saturated* and a Hamiltonian cycle will be formed.

Often, the proposed link is not possible as one of the lattice points is already

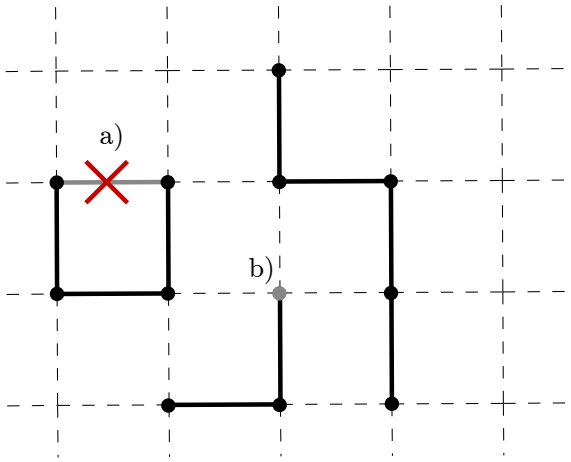


Figure 5.3: Situations forbidden when generating the lattice walks. a) shows a link which would create a subcycle and b) shows, highlighted in grey, a point which is in a dead end.

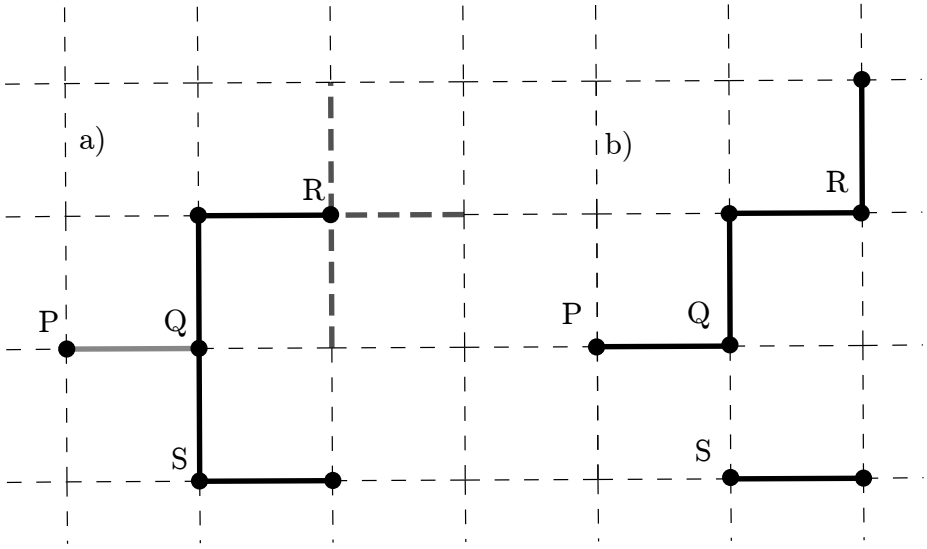


Figure 5.4: The procedure for two-matching. The initial situation is shown in a) where the proposed link between P and Q is in solid grey. Possible extensions to R are highlighted in dashed grey. A possible final situation is shown in b), where a link has been add from R upwards, P and Q have been linked and Q and S have been unlinked.

saturated. In these cases a procedure called *two-matching* is used. Assume the proposed link is between points P and Q as shown in Fig 5.4 a). P is unsaturated, but Q is already linked to two other points. In two-matching, one of the current links to Q is followed, chosen at random, until the terminus of that sub walk is found, called R. A list is then made of the possible links that can be made from R which do not result in dead ends or subcycles, as shown by the dashed grey lines in a). If there are possible links to be made from R, one is made at random. Then, the link from Q to S, the first vertex in the opposite direction from R, is broken and the link between P and Q made on the condition that P is still unsaturated after R has been linked and that linking P and Q does not produce a dead end or subcycle. Fig 5.4 b) shows a possible situation after two matching.

Hamiltonian walks created in this way are not quite sampled uniformly from all possible Hamiltonian walks, but the bias is small, as demonstrated by Lua et al. [22]. To obtain walks of any possible length, we take a random subwalk from the resulting Hamiltonian walk.

5.1.2 Off-lattice walks

We also look at off-lattice walks, in particular ideal chains which have been used as models of flexible polymers under theta conditions. We generate these step-by-step using *Markov chain* processes, where each step knows only where it will step from with no knowledge of the rest of the walk [1]. Examples of the walks can be seen in Fig 5.5. For the simple unconfined case we choose a point in $\mathbb{R}^3$ to begin the walk, typically the origin although this can always be changed by translation if necessary, and choose a point on the unit sphere centred on this point. We add this new point to the walk and choose the next point in the same manner, but with the unit sphere now centred on the new point. This guarantees each step is the same length also. A uniform scaling transformation can be used to adjust step length, but like the origin position this does not fundamentally affect walk properties.

Spherically confined walks

A particularly relevant shape of confinement is the sphere, which mimics conditions in viral capsids or in lipid micelles, for example. A naive approach to generating confined ideal chains is to generate the walk in the same manner

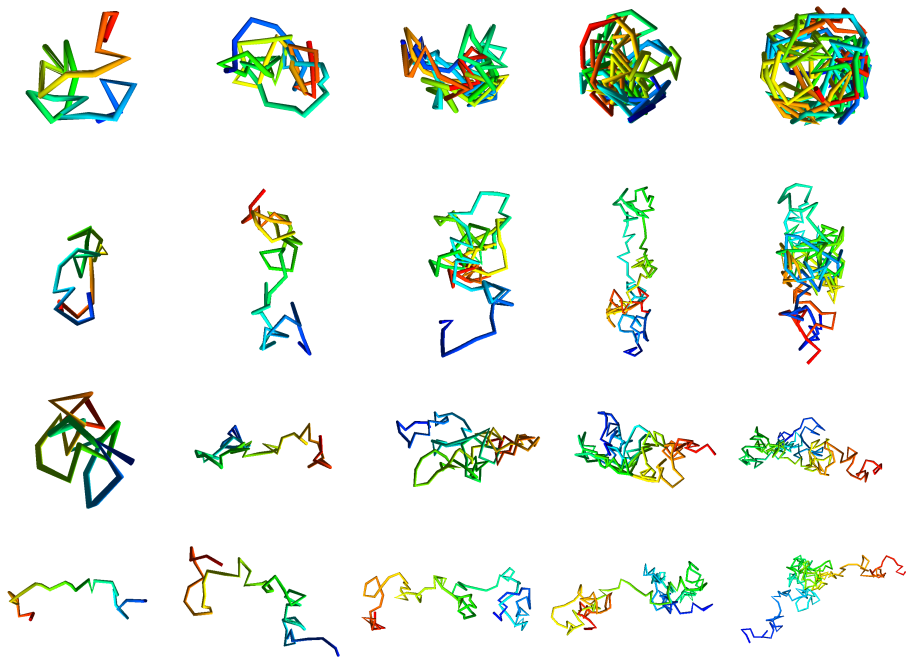


Figure 5.5: Examples of random ideal chains. From top to bottom the confining volumes are: spheres, tubes, slits and no confinement. From left to right the lengths are: 25, 50, 100, 150 and 200.

as for unconfined chains, but if a step were to take the walk beyond the confinement, it is rejected and another step is sampled, until a step within the confinement is found. The boundary of such a walk is said to be *absorbing*. The problem with this is that the vertices of such a walk are not uniformly distributed in the confining volume. In fact, the vertices of such walks are less likely to be within a step length of the boundary than would be expected [149]. One way of realising this is to consider the places a given point in the volume may be reached from. Fig 5.6 shows two points in a circle, and the locations they may be reached from. The point a) may be reached by points on the unit circle surrounding it. This is partly true for point b), except the points beyond the boundary will never be reached by the walk and so b) cannot be reached from them. It is easy to see that there are more ways to reach a) than b) and so we expect a) to be visited with higher probability. Hence, we expect fewer walk vertices within a unit of the boundary, and for this to be more apparent the closer to the boundary one gets as more of the unit circle lies outside the boundary.

It would be good to have a Markov chain process to generate confined off-lattice walks if possible. Happily, Diao, Ernst, Montemayor and Ziegler [149]

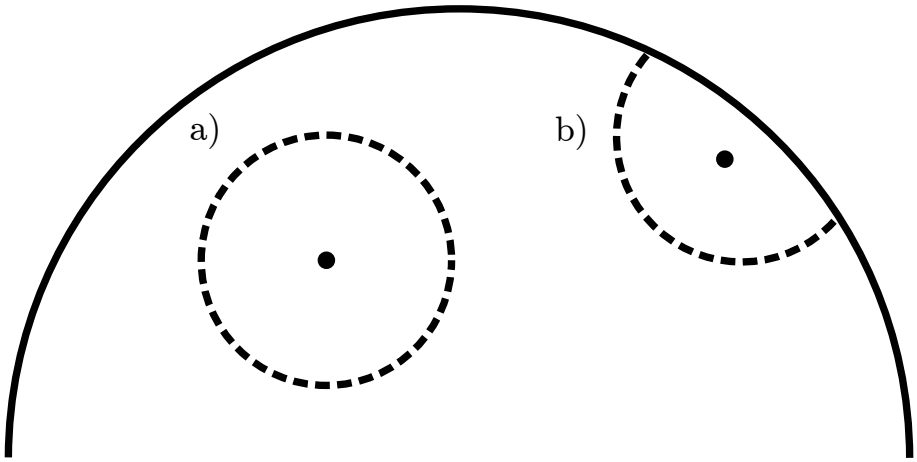


Figure 5.6: Two possible points a random walk in the circle may visit. a) shows a point away from the boundary, and b) shows a point near the boundary. The dashed lines show the locations from which the point may be reached by a walk.

devised a solution for this for spherically confined walks. While the vertices are not exactly uniformly distributed, the approximation is very good. While a random walk is possible in a sphere of radius of half a step length, this algorithm requires a radius of at least one whole step length. The procedure they devise and we use is as follows.

The first point of the walk, X is sampled uniformly from the confining sphere of radius R, centred at the origin. There are a number of ways to do this, but the very simple way we use is to sample x, y and z uniformly in the range $[-R, R]$, and then check if this point lies inside the sphere. If not, then the point is rejected and another point sampled until a valid coordinate is found. Then, if the distance from X to the origin, r, is less than $R-1$, the next point is sampled uniformly from the unit sphere centred at X just as for the unconfined walks. In other words, if X is further than a unit from the boundary, we proceed as for the unconfined walks. If $r \geq R-1$, or the point is within a unit of the boundary, then special measures must be taken.

To choose the next point, two angles are chosen with respect to the radial line of the boundary sphere which passes through X. The angle around the radial direction ϕ is simply chosen uniformly in the range $[0, 2\pi)$. The angle from the outward radial direction, θ ordinarily could be sampled by choosing $-\cos\theta$ uniformly in $[-1, 1]$. This would be appropriate for sampling a point on the sphere uniformly, but of course we don't want to do this.

We have two conditions we wish to satisfy when choosing θ. First, we don't want to choose a θ which takes the walk outside the boundary sphere. Second, we want to sample angles which are closer to the boundary more frequently, to offset the bias of the walk to avoid the boundary if we sample across angles uniformly. An exact solution to this problem to give uniform walk vertex distribution in the sphere is not known, but Diao et al. make the following approximation.

The range of $-\cos\theta$ is split into three distinct categories: angles which, on a unit sphere about X, lie further than a unit from the boundary, angles within a unit of the boundary, and outside the boundary. The values of $-\cos\theta$ which lie within these categories are:

$$-\cos\theta = \begin{cases} \text{further than a unit from boundary,} & -1 \leq -\cos\theta \leq a; \\ \text{within a unit of boundary,} & a < -\cos\theta \leq b; \\ \text{outside boundary,} & b < -\cos\theta \end{cases} \tag{5.1}$$

where

$$a(r,R) = \begin{cases} \frac{R^2-r^2-2R}{2r}, & r > R-1 \text{ and } R-1 > \text{Min}(r,|r-1|); \\ -1, & r > R-1 \text{ and } R-1 \leq \text{Min}(r,|r-1|); \\ 1, & r \leq R-1 \end{cases} \tag{5.2}$$

and

$$b(r,R) = \begin{cases} \frac{R^2-r^2-1}{2r}, & r > R-1; \\ 1, & r \leq R-1 \end{cases} \tag{5.3}$$

These values are illustrated in Fig 5.7.

$-\cos\theta$ is then sampled differently in each range. Outside the boundary, the probability of sampling is zero. Well within the boundary, the probability of being sampled is uniform, just as the unconfined case. Within a unit of the boundary, the approximation is to sample with a linearly increasing probability the closer to the boundary the angle is. This gives us the probability density function (PDF):

$$\text{PDF}(-\cos\theta) = \begin{cases} \frac{1}{2}, & -1 \leq -\cos\theta \leq a; \\ \frac{1}{2}(1 + c(-\cos\theta - a)), & a < -\cos\theta \leq b; \\ 0, & b < -\cos\theta \end{cases} \tag{5.4}$$

where

$$c = \frac{4r((r+1)^2 - R^2)}{(2R-1)^2} \tag{5.5}$$

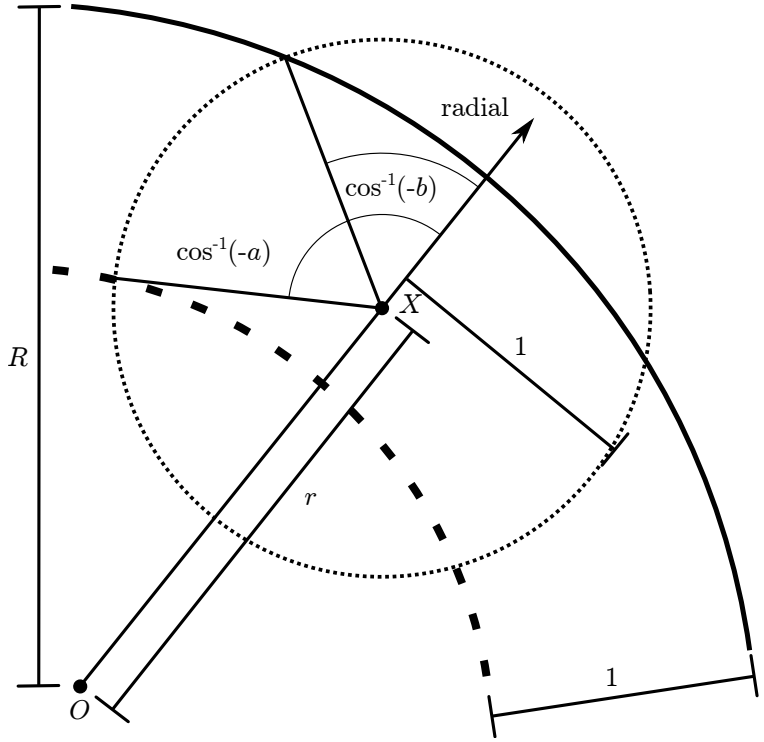


Figure 5.7: Diagram showing the angles used in generating uniformly distributed vertices for random walks in the sphere. The current walk vertex position is X, the next step of the walk will be chosen from the unit sphere surrounding this point. The walk is confined to the sphere centred at O with radius R, and X is a distance r from O.

for normalisation. The resulting PDF for various r in a sphere of $R = 3$ is given in Fig 5.8.

The actual radial distance distribution of walk vertices is given in Fig 5.9. This was generated from the end-points of 10 million walks of ten steps long for both the absorbing boundary method, and the method just described. As can be seen, the method of Diao et al. is a great improvement on the absorbing boundary, getting much closer to the distribution of uniform points in the sphere.

Walks confined to a slit

While the method of Diao et al. was originally designed for walks confined to the sphere, we would like to investigate walks in other shapes of confinement also. These walks will face a similar problem if an absorbing boundary is used, and so some modification of the method just outlined needs to be used to ensure a uniform distribution of vertices.

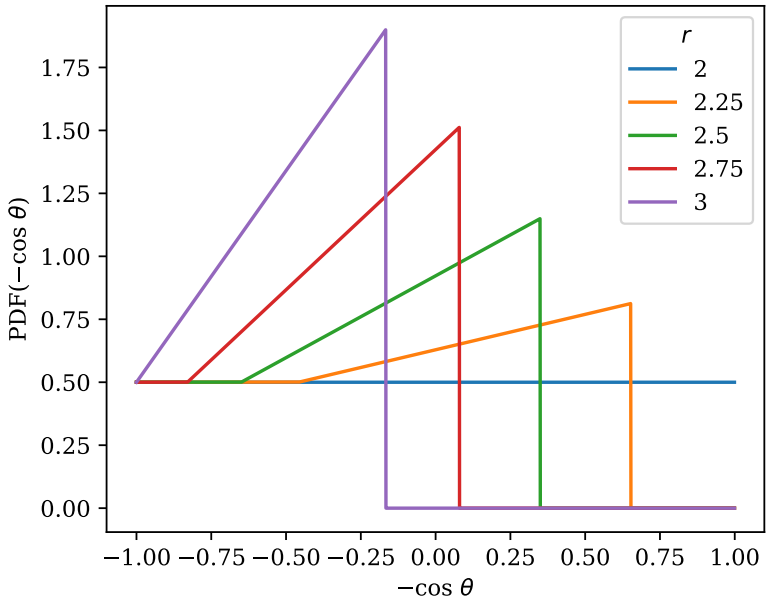


Figure 5.8: The PDF from which $-\cos\theta$ is sampled according to [149] for walks in a sphere of radius 3. The radial distance of the point to be stepped from, r, is given by different lines. Values to the left correspond to angles further within the boundary sphere, and values to the right to angles towards and outside the boundary sphere. The point where the PDF begins to rise is at a, and the point where it drops to zero is at b as in Fig 5.7.

A confinement shape commonly explored in the polymers and random walk literature is that of two parallel planes, also called a slit, which has applications to thin film solutions and microfluidic environments [102, 103, 104, 105]. This is a relatively straightforward case to extend the method of Diao et al. to. Taking advantage of differing length scales, we can simply model each confining plane as if it is a very large sphere, much larger than the step length of the walk. This will present an approximately flat surface when picking angles for each step.

If the point to be stepped from, X, is further than a unit from each plane, then points are picked uniformly on the unit sphere centred on X as before. If X is within a unit of a plane, we use the modified Diao et al. approach to choose the angles. Say the radius of the large sphere we are using to approximate the plane is R. In practice we take $R = 10,000$. The radial distance from the centre to X as used in the original method is taken to be R minus the distance from X to the boundary. The radial line from which the angles are measured is taken to be the normal to the plane which passes through X and is parallel to the slits. With these modifications, an angle can now be sampled with the approximately

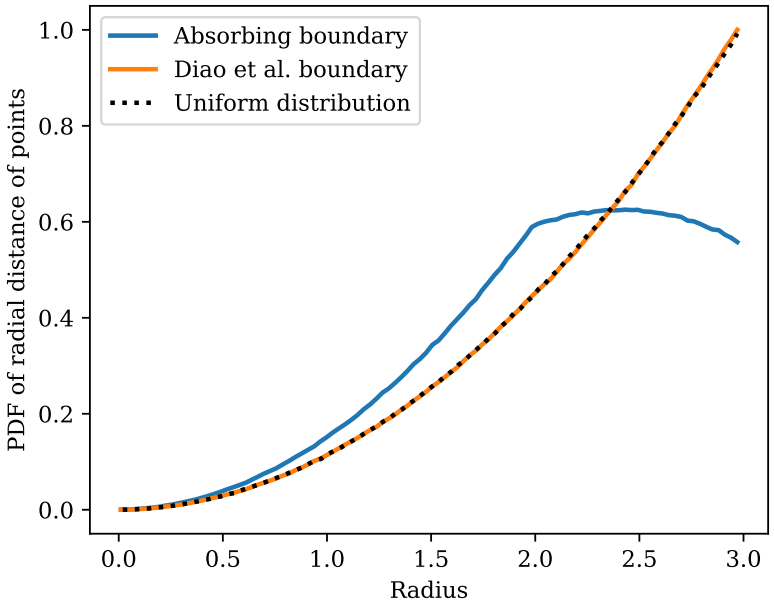


Figure 5.9: The distribution of radial distances of random walk points in a sphere, compared to uniform points. Shown are walks created with an absorbing boundary, and with the correction by Diao et al. [149]. The plot is a histogram of data, with 10 million points for each line.

uniform distribution seen for the sphere.

The restriction of this modification is that the planes cannot be taken too close together. If the unit sphere around X goes beyond both planes, then this must be taken account of when choosing angles. To avoid this, the planes can be kept at least two units apart. This way X only ever lies within a unit of one plane, and never both.

Walks confined to a tube

Another highly relevant confinement shape with application to microfluidic channels, pores and other capillaries is that of cylinders or tubes [99, 100, 101]. Again, we need a Diao-esque procedure to obtain a uniform vertex distribution. Unfortunately, this cannot be as easily achieved as with the walks in slits. We have however found a workaround which modifies the approach in the sphere to produce an almost uniform distribution. The approximation is poorer than in the sphere, but much better than an absorbing boundary.

Say the axis of the tube lies on the z-axis and the tube extends infinitely in z. The tube has radius R. We are stepping from a point, X, to a point on

the unit sphere centred on X. Once again, if X is further than a unit from the boundary then we choose a point on the unit sphere uniformly. Within a unit of the boundary, we must account for the fact that a rejection-sampled point will avoid the boundary.

Here is where the difference between the cylinder and the sphere becomes important. One cannot choose a polar angle and an azimuthal angle about the radial line through X as in spherical or slit confinement, as the intersection of the unit sphere about X and the boundary of the tube is not circular. This means that depending on the azimuthal angle chosen, the range of polar angles which lie within the boundary will change, which surely will have consequences for the PDF used.

To avoid this complexity, we orient the unit sphere we step to not along the radial line through X, but parallel to the z-axis. We choose a ϕ now from the positive z direction, taking $\cos\phi$ uniformly in the range $[-1,1]$ as ϕ here is our polar angle. This is shown in Fig 5.10 a). Having decided the z-coordinate of our random walk step, we now must find the xy coordinate. As z is fixed, we have a cross-section of the boundary tube which is normal to the tube axis, ensuring that the shape of the boundary cross-section is always circular. We choose a θ to point to the slice of unit sphere at the given z, as shown in Fig 5.10 b). This slice may extend beyond the boundary, but we can use a Diao-like approach to deal with this.

As before, we choose our θ to be from the outer radial direction. In the spherical boundary case, our PDF for this was in terms of $-\cos\theta$, as θ there was a polar angle. Here, it is an azimuthal angle, so our PDF is in terms of θ instead. Also, in the spherical case θ only fell within $[0,\pi]$, as ϕ fell within $[0,2\pi)$, covering the entire sphere. As we would like to use the methodology of Diao et al. as closely as possible, we still choose a θ between $[-\pi,0]$ initially, and then randomly choose whether it is measured clockwise or anticlockwise from the outer radial direction as the problem is symmetric about this line. In this way we cover angles across the entire sphere.

In the method of Diao et al. the values of r (the radial distance of X) and R (the radius of the confining volume) are used to determine the angles of θ which point to a unit from the boundary, and to the boundary itself. The quantities a and b capture these angles. The quantity c also depends on r and R for correct normalisation of the PDF. However, the assumption in the original work was that θ was an angle on a unit circle. Here, we are taking slices of the unit sphere

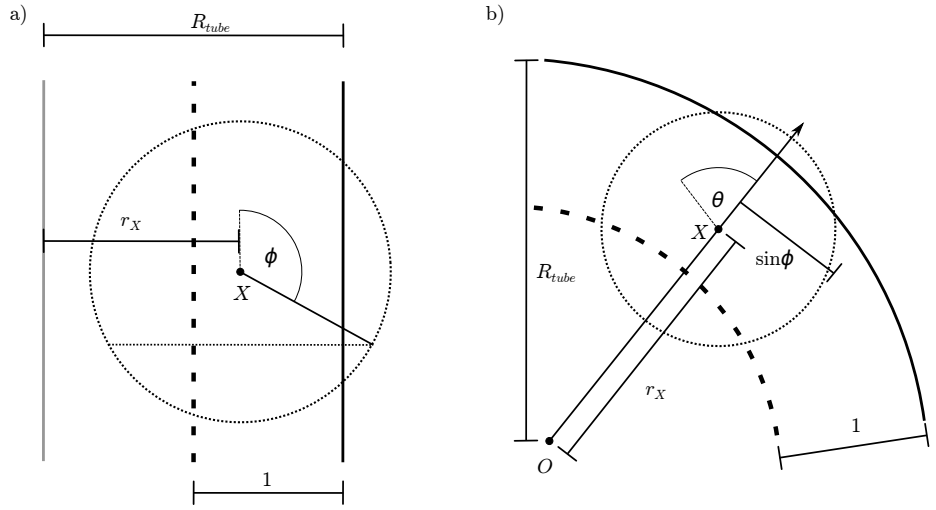


Figure 5.10: a) Diagram shows the unit sphere around the point X, and the cylinder of radius R_{tube}. The vertical dashed line is a unit from the boundary. Drawn also is the angle ϕ, with the horizontal dashed line giving the slice of the unit circle corresponding to this. b) Diagram shows the z-slice of a) at a given ϕ. Note that the radius of the slice of the unit sphere is $\sin\phi$ and so is smaller than 1. Indicated also is the angle θ which here is being measured anticlockwise from the radial line. It is uniform random whether we measure θ clockwise or anticlockwise.

of constant z. This means the radius of the circular slice depends on ϕ, our polar angle. To be exact, for the unit circle, the radius of the circular slice is $\sin\phi$, as indicated in Fig 5.10 b). To account for this discrepancy, we scale the entire z-slice, so that the effective radius of our unit-circle slice is 1. This gives an effective r of

$$r_{\text{eff}} = r_X / \sin\phi \tag{5.6}$$

where r_X is the radial distance of X, and an effective R of

$$R_{\text{eff}} = R_{\text{tube}} / \sin\phi \tag{5.7}$$

where R_{tube} is the actual radius of the tube. These effective radii are then used to calculate a, b and c.

The essential property we wish to retain from the approach of Diao et al. is to account for the boundary avoiding properties of the absorbing walks. To do this, we want the PDF from which we sample θ to be uniform for angles pointing more than a unit from the boundary, increasing according to some function for angles within a unit of the boundary, and then zero outside of the boundary. However, if we use the same linear increase of probability within a unit of the boundary, we see that we overcompensate and push the walk to the

boundary. One might expect this as for a fixed r and R, there is less of the unit sphere extended beyond the boundary of the cylinder than the sphere, resulting in less effect to compensate for.

To account for this, we need a PDF(θ) which does increase towards the boundary, but not as steeply near the boundary as a linear rise. In trying to understand which form this should take, we corresponded directly with Diao in December 2017 as to why a linear rise was chosen originally. He said, 'The linear growth was the first natural candidate and perhaps it was just pure luck that it worked. Given how well the simulation went, it is quite plausible that this indeed gives a uniform distribution for the vertices, but we were not able to prove it.' Given this, we felt liberated to experiment with different forms to find something that worked 'well enough'.

The general form we use is for the probability within a unit of the boundary to grow not as θ, but as θ^n, where $n < 1$. The resulting PDF with this modification takes the form:

$$\text{PDF}(\theta) = \begin{cases} \frac{1}{\pi}, & -\pi \leq \theta \leq \alpha; \\ \frac{1}{\pi}(1 + c(\theta - \alpha)^n), & \alpha < \theta \leq \beta; \\ 0, & \beta < \theta \end{cases} \tag{5.8}$$

where

$$\alpha = \cos^{-1}(-a) \quad \text{and} \quad \beta = \cos^{-1}(-b) \tag{5.9}$$

which are the angles equivalent to the boundaries used in the sphere case.

To calculate the necessary normalisation factor, c, we demand that the integral of the PDF equal 1. If we split the PDF into the areas A and B as indicated in Fig 5.11, we can simplify our calculation. We can easily see that:

$$A = \frac{1}{\pi}(\beta + \pi) = \frac{\beta}{\pi} - 1 \tag{5.10}$$

and we know that:

$$A + B = 1 \implies B = 1 - A = -\frac{\beta}{\pi} \tag{5.11}$$

so we calculate:

$$\begin{aligned} B &= \int_{\alpha}^{\beta} \frac{c}{\pi}(\theta - \alpha)^n d\theta \\ &= \frac{c}{\pi}\left(\frac{(\theta - \alpha)^{n+1}}{(n+1)}\right)\Bigg|_{\alpha}^{\beta} \\ &= \frac{c(\beta - \alpha)^{n+1}}{\pi(n+1)} \end{aligned}$$

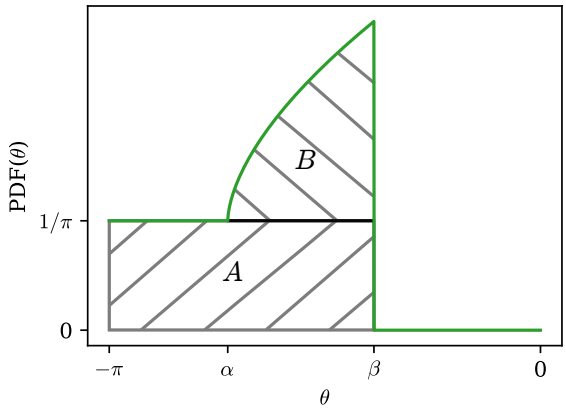


Figure 5.11: Diagram showing the PDF of θ in the cylinder, where the probability grows as θ^n within a unit of the boundary and $n < 1$. Areas A and B are considered when calculating the normalisation factor c.

rearranging for c and substituting in $B = -\beta/\pi$ we get:

$$c = \frac{-\beta(n+1)}{(\beta-\alpha)^{(n+1)}} \tag{5.12}$$

We numerically experimented with n in the range $[0.5, 0.7]$ to find a value which reduced deviation from uniform point distribution as much as possible. We reached a value of $n = 0.6$ as giving the closest to uniform distribution. Before going into detail as to how this value was reached, it will give some perspective if we show the end result first. Fig 5.12 shows the form of the PDF, and Fig 5.13 shows a comparison of the radial distribution of random walk end-points in the tube using no correction, the original correction of Diao et al. and our modified PDF. As can be seen, both the linear in θ correction and the $\theta^{0.6}$ correction do much better than an absorbing boundary, but $\theta^{0.6}$ provides a significant improvement over the original correction.

Now, to see how we came to a value of $n = 0.6$ we look at the difference of the radial distribution from uniform for various values of n in the range $[0.5, 0.7]$ as plotted in Fig 5.14. The first thing to note is that no value performs quite as well as the original correction performed in the sphere. However, the deviations are not vastly larger in scale. It is clear that certain values of n perform better within different distances from the boundary, so we seek a value which gives the best compromise and is not very wrong anywhere, but also minimises the total deviation from uniform. The plots shown in Fig 5.15 give data showing a) the total deviation from uniform and b) the maximum deviation from uniform.

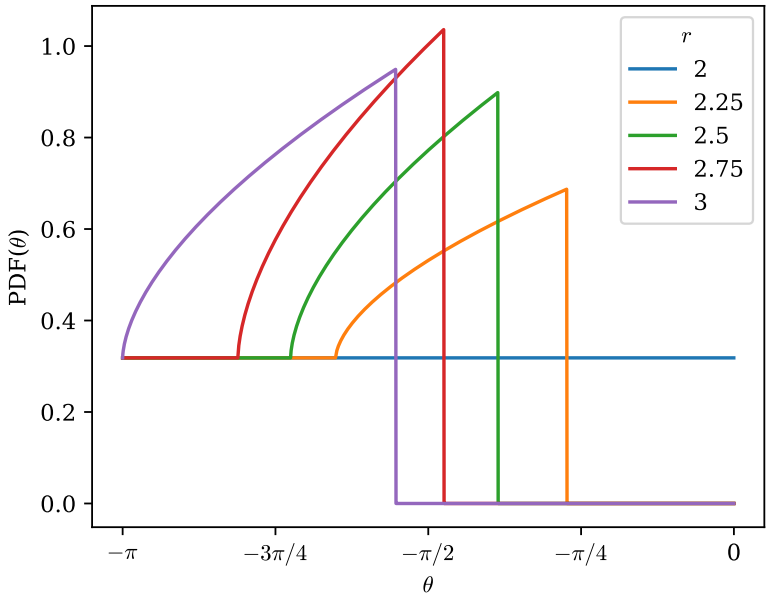


Figure 5.12: The PDF from which $-\theta$ is sampled for walks in a tube of (effective) radius 3. The (effective) radial distance of the point to be stepped from, r, is given by different lines. Values to the left correspond to angles further within the boundary cylinder, and values to the right to angles towards and outside the boundary cylinder. Within a unit of the boundary, $PDF(-\theta)$ rises as $\theta^{0.6}$.

From this, we find a value of $n = 0.6$ to provide a good compromise. However, the scale of the deviations seen from all the values is relatively small and one would hope that the effect on the knotting of this ensemble would be minimal. To ensure this is the case, while we take $n = 0.6$ to be our best value, we have duplicated all the knotting analysis about to be presented for walks with a value of $n = 0.5$ also.

It should be said here that the method presented for generating random walks in the tube is almost certainly not the 'correct' way. What we hoped to achieve was an approach which gave approximately uniform vertex distribution in a tubular confinement and the data given here shows that the deviation is minimal, especially when compared to an absorbing boundary. While it would be good to have a more rigorous solution to this problem we thought that the impact on knot statistics of these walks was likely to be minimal and the results we do get will hopefully be instructive as a comparison for later improvements to the method.

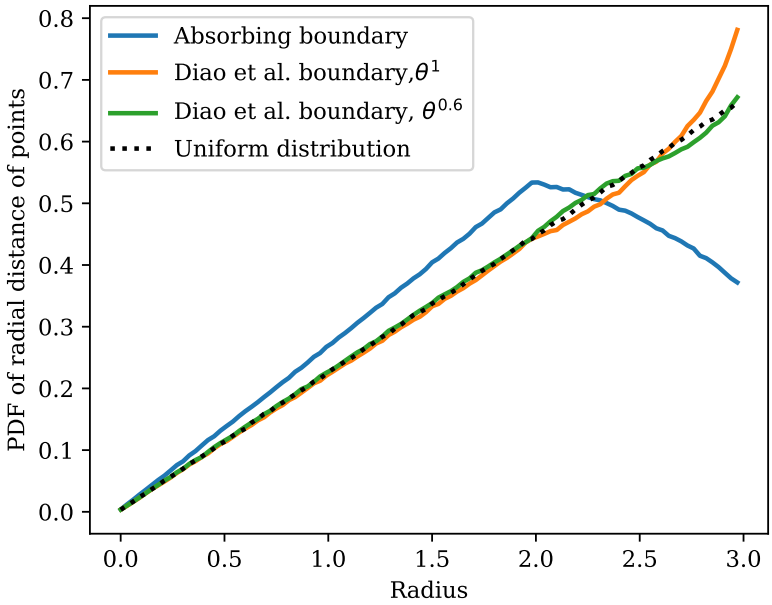


Figure 5.13: The distribution of radial distances of random walk points in a tube, compared to uniform points. Shown are walks created with an absorbing boundary, and with the original correction by Diao et al. [149], and our modified form where the PDF rises as $\theta^{0.6}$ as opposed to θ. Plotted is histogrammed data, with 10 million points for each line.

5.2 Knotting in confined random walks

Having covered the details of how our walks are generated, we can now ask questions of their knotting. We begin by asking how the knotting between lattice and off-lattice walks compare and if there are notable differences under sphere and virtual closure. We will pay particular attention to their weak knotting as a feature of knotting unique to open curves. Then, we will do a wider ranging survey of knotting off-lattice, starting with just spherically confined walks. By varying the size of the confining sphere we probe how the degree of confinement of these walks affects the knot statistics. Finally we look at off-lattice walks in different shapes of confinement, reducing the number of confined dimensions from three in the sphere, to two in the tube and one in the slit, to determine the knotting behaviour of walks in these physically relevant geometries.

5.2.1 Comparing lattice and off-lattice walks

To begin we compare knotting on and off-lattice. We confine the lattice walks to a cube, and investigate off-lattice walks both in the sphere and unconfined.

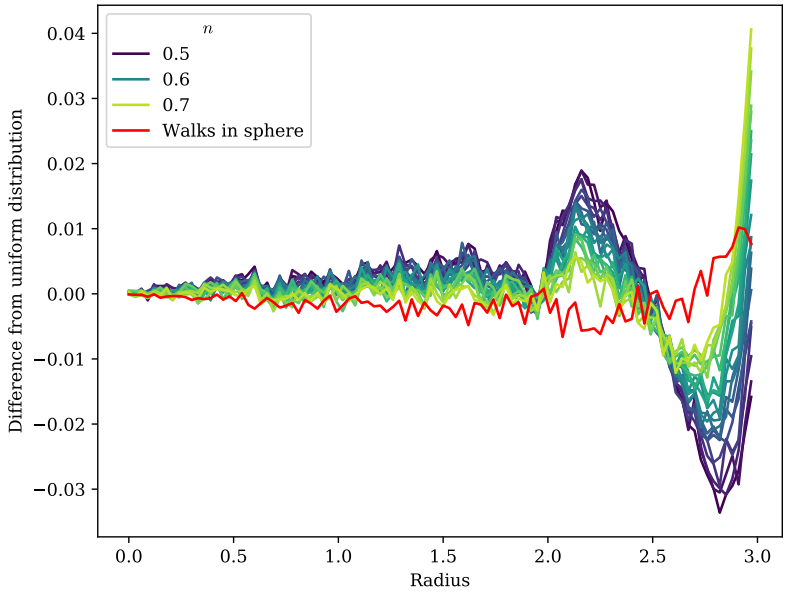


Figure 5.14: The difference of radial distribution of random walk end-points in the cylinder using a Diao-style boundary where probability grows as θ^n. Shown in red also is the difference from uniform for the original corrected walks in spherical boundary conditions. Plotted is histogrammed data, with 10 million points for each line.

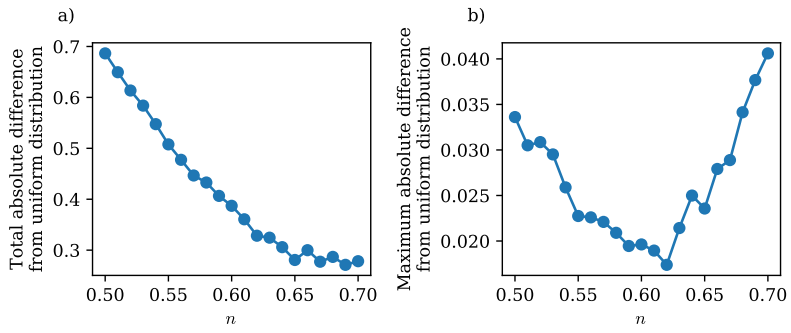


Figure 5.15: a) The sum absolute difference from uniform distribution of the data shown in Fig 5.14. b) The maximum difference of the same data.

We use two different sized lattices: $6 \times 6 \times 6$ and $7 \times 7 \times 7$. To provide the most comparable confinement in the off-lattice case, we choose sphere radii which give an equal volume to our two lattice cubes. This is given by:

$$R^3 = \frac{3}{4\pi}(L-1)^3 \tag{5.13}$$

where R is the radius of equivalent volume to the lattice of $L \times L \times L$ nodes. Step length is assumed to be 1 in all cases. This gives us sphere radii of approximately 3.1 and 3.72 respectively.

For each walk model and confinement shape we investigate the knotting of a variety of lengths, limited by the saturation length on lattice. For each set of parameters we generate an ensemble of 10,000 walks. Each walk is analysed using 100 sphere closures and 100 virtual closures as outlined in Chapter 3. This is the same as was done for proteins in Chapter 4.

Fig 5.16 a) shows how the probability of knotting varies with walk length, where being knotted means 50% or more closures are non-trivial knots. The solid lines are knotting analysed by virtual closure, and the dashed lines are using sphere closure. Clearly, the longer the walks are the more likely they are to be knotted by any measure. The confined off-lattice walks are more likely to be knotted than the unconfined off-lattice walks, but the reduced flexibility of the lattice walks means they are less likely to be knotted than even unconfined off-lattice walks, at these length scales. We also see that the tighter the confinement is, the more likely a walk of a given length is to be knotted. All of this is just confirming what was known already from the literature.

The comparison between virtual and sphere closure here is interesting. For the unconfined walks, we see that the methods agree very closely on the probability of knotting. For confined off-lattice walks, virtual closure gives a slightly higher probability of knotting, although the proportional difference in this reduces at longer lengths, where essentially every walk is knotted regardless of the detection method used. The biggest discrepancy is for the confined lattice walks where the probability of knotting under virtual closure can be twice as high as under sphere closure. We expect the knots seen on lattice to be less complex than those off-lattice and so it is more likely here that discrepancies in knot classification are between simple knots and unknots. Off-lattice we expect each closure method to disagree on exactly which knots are present, but with neither showing much unknotting.

Fig 5.16 b) then shows the probability of knotted walks being weakly knotted,

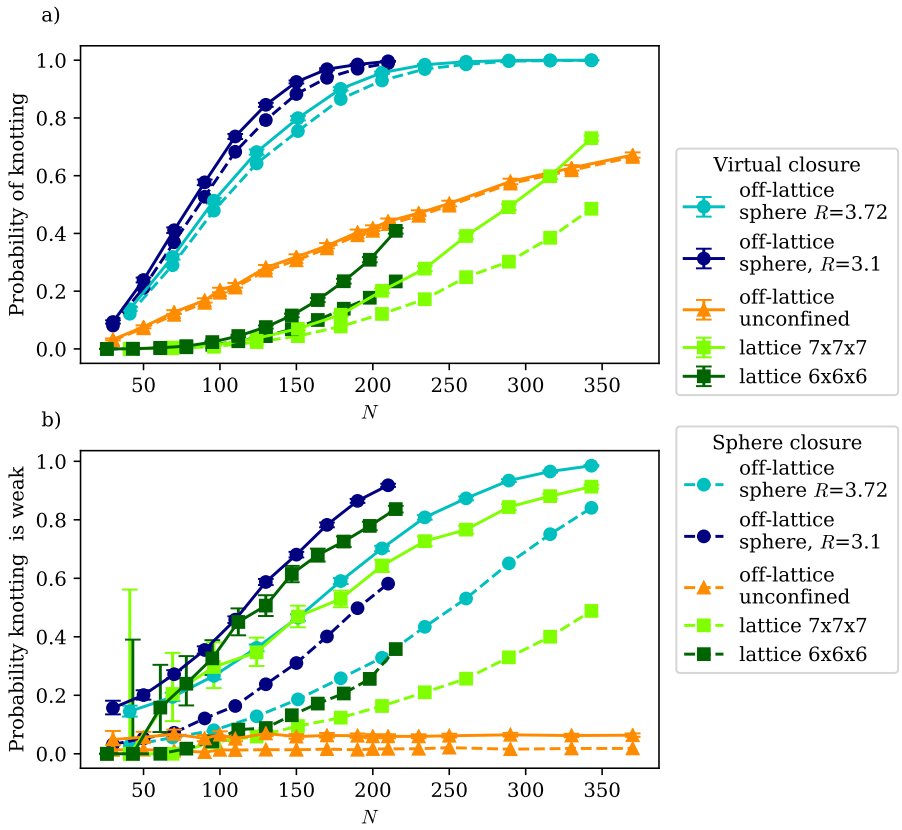


Figure 5.16: a) The fraction of random walks which are knotted under virtual closure (solid line) and sphere closure (dashed line). b) The fraction of knotted random walks which are weakly knotted under each method as in a). Off-lattice walks confined to spheres have circular markers, on-lattice walks confined to cubes have square markers, and off-lattice unconfined walks have triangular markers.

where weakly knotted means the most common knot appears in fewer than 50% of closures. Again, solid lines are weakly knotted under virtual closure, and the dashed lines are using sphere closure. At these length scales we see essentially no weak knotting in unconfined walks, as was hinted at by Millett et al. [26]. Remarkably we see very similar behaviour for both confined walk models. We would expect to see a greater proportion of weakly knotted walks in confinement, especially as the walks get longer and are affected by the confinement more [28]. Indeed this is what we see, but it is striking how closely aligned the probabilities are under virtual closure for on and off-lattice walks in equivalent volumes.

We would expect to see a smaller proportion of weak knots under sphere closure and this is the case here. As fewer knot types are accessible with sphere

closure than virtual closure, we see it is more likely for a single knot type to dominate. This is more apparent for confined lattice walks than for the confined off-lattice walks. As the lattice walks are simpler space curves, there are fewer distinct knot types possible compared to the infinitely flexible off-lattice walks. This reduced competition between knot types leads to fewer weak knots on-lattice. What is remarkable is that the extra competition added by shifting to virtual closure brings the two models so close together.

Fig 5.17 a) and b) show the same data as Fig 5.16 but plotted instead from the point of view of unknotting and strong knotting, with a log y-axis. The main point of this is to show that the unconfined walks, which here have been extended to longer lengths, do indeed show the expected exponential decay of unknotting with length. This trend is not seen in the confined walks where the additional pressure towards knotting given by the confinement encourages a sharper than exponential decay of unknotting and strong knotting. The extra unconfined walk data also shows that, at these length scales at least, there is no appreciable variation in weak knotting with length.

5.2.2 Spherically confined off-lattice walks

In comparing different walk models in the previous section, we kept confinement size fixed and varied length. This has the unfortunate effect of conflating knotting due to length and knotting due to confinement. In order to investigate these effects separately, we now look solely at off-lattice walks, which gives us finer control over the parameters varied. We will keep the shape of the confinement spherical for now. The results presented will also be from virtual closure. The differences between virtual closure and sphere closure seen in the plots already shown can be assumed to be true here also.

We plot how spherical confinement radius affects knotting probability in Fig 5.18 a). Each marker shape corresponds to a single length of walk and so the effect of the confinement is isolated. There is a clear trend towards increased knotting as confinement becomes smaller. This knotting can be seen to saturate at 100% in the longest walks. Longer length increases knot probability at all levels of confinement.

Looking to weak knotting, Fig 5.18 b) shows the probability of knotting being weak for the same walks. Similarly, we see that tightening confinement increases the proportion of knots which are weak, and again this saturates at 100% for the longest walks in the tightest confinements. As we saw when we looked at

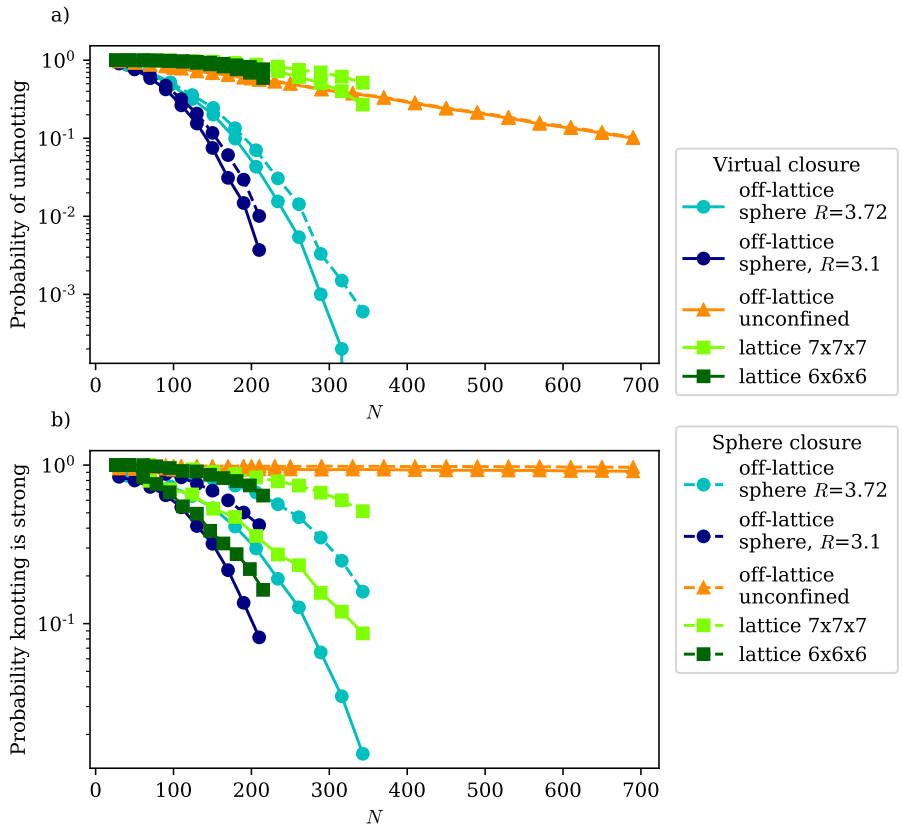


Figure 5.17: a) The fraction of random walks which are unknotted under virtual closure (solid line) and sphere closure (dashed line). b) The fraction of knotted random walks which are strongly knotted under each method as in a). The same data as Fig 5.16, but here with additional unconfined walk lengths.

unconfined walks previously, length does not make a difference here to weak knot proportion without confinement. This means that all the walks start at the same very low weak knotted proportion, and the rise to 100% weak knotting is faster the longer the walks are.

This suggests that this is not the most natural way to present this data. Clearly, the degree to which the walks are confined is different for each length. To this end, we plot the same walk data, but this time we plot knot/weak knot probabilities against the *degree of confinement*. We define the degree of confinement of walks with length N and in sphere of radius R as:

$$\text{degree of confinement} = \frac{\langle R_g(N, R=\infty)\rangle}{\langle R_g(N, R)\rangle} \tag{5.14}$$

where $\langle R_g(N,R)\rangle$ is the average radius of gyration of the confined walks, and $\langle R_g(N, R=\infty)\rangle$ is the average radius of gyration of unconfined walks of the same

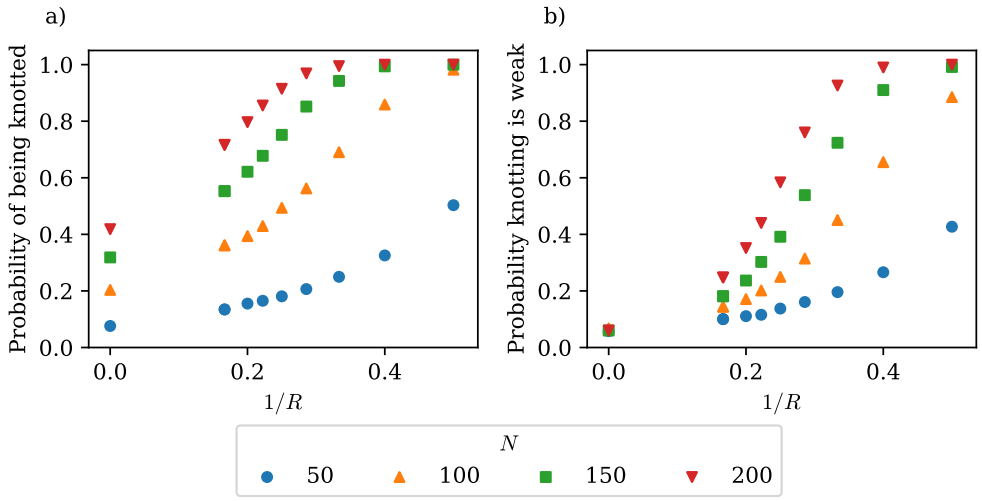


Figure 5.18: a) The fraction of spherically confined off-lattice random walks which are knotted under virtual closure as confinement radius R varies. b) The fraction of the knotted random walks which are weakly knotted under virtual closure. Data in both plots plotted against $1/R$. Unconfined walks have $1/R = 0$. Error bars are no larger than the markers.

length. Larger values of degree of confinement correspond to walks which are relatively more confined, and a value of 1 corresponds to unconfined walks.

We plot the same data as we just analysed now against the degree of confinement in Fig 5.19. Looking at the probability of knotting in a), clearly as the degree of confinement increases, knots become more likely. It is also clear that the length of the walks plays an additional part in their knotting, which we expect as this is true even for unconfined walks.

The real interest here is in the probability of the knotting being weak, as shown in Fig 5.19 b). All of our data has now collapsed onto essentially the same curve here, with greater degree of confinement giving greater probability that knotting is weak. There are small differences with length, with longer walks having a slightly higher tendency to weakly knot at the same degree of confinement, but largely the degree of confinement is the most important factor for weak knotting.

It would seem reasonable that as the fraction of knotted walks which are weakly knotted increases the coverage of the most common knot over all closures should decrease, on average. We plot the ensemble average coverage of the most common knot in Fig 5.20 and see that indeed this is the case. Strikingly, we see that by plotting against the degree of confinement in b), all the data falls

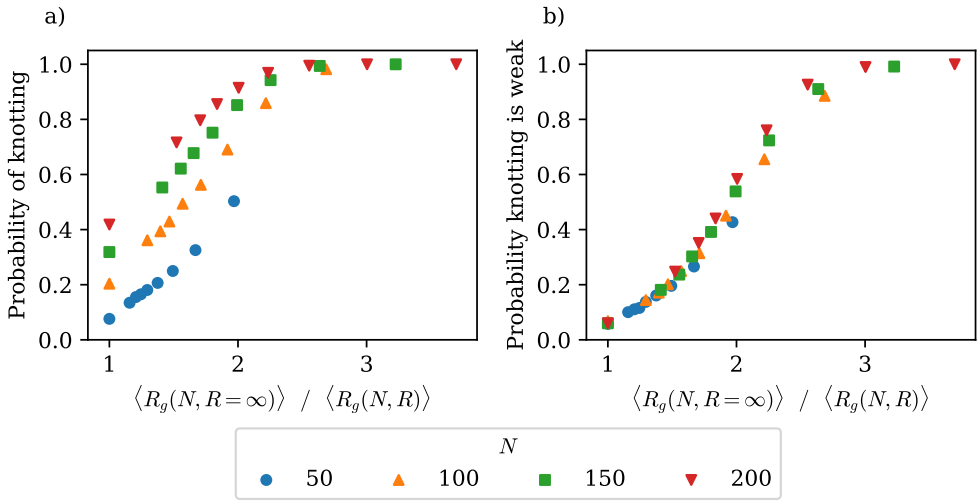


Figure 5.19: a) The fraction of spherically confined off-lattice random walks which are knotted under virtual closure as degree of confinement varies. b) The fraction of the knotted random walks which are weakly knotted under virtual closure. Same data as Fig 5.18. Again, error bars are on the order of the marker size or smaller.

onto the same curve again. As the degree of confinement increases, the average coverage of the most common knot decreases i.e. the knots get weaker. At the highest degrees of confinement investigated, we are actually limited in accuracy by only taking 100 closures, as essentially every closure here gives a different knot type. Again, we do see a slight length dependence, but this effect is small.

We can take a look at this in more detail by binning all the walks into ranges of degree of confinement. For each set of walks, we then plot the distribution of the coverage of the most common knot in Fig 5.21. It is obvious where the averages we've just looked at originate. The distribution for walks with a low degree of confinement peaks at a high fractional coverage. This peak then shifts lower as the degree of confinement increases. When we choose to include walks whose most common knot is the unknot, a), we see a relatively sharp peak at the highest fraction of closures for the least confined walks. Excluding walks whose most common knot is the unknot, b), we see that this peak is much broader, indicating that the unknotted walks have the least ambiguity in knot type. This effect is much less noticeable in the more confined walks where unknotting becomes much rarer.

We have hypothesised before that weak knotting is related to end-point position. The more buried the end-point is within the bulk of a space curve,

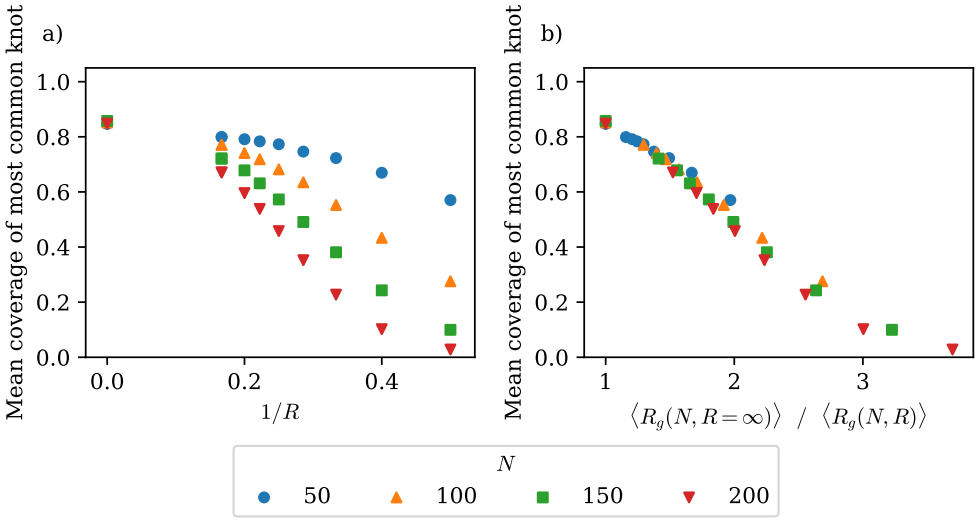


Figure 5.20: The ensemble average coverage of the most common knot in spherically confined off-lattice walks plotted a) against inverse confinement radius and b) degree of confinement. Only walks whose most common knot is non-trivial are considered.

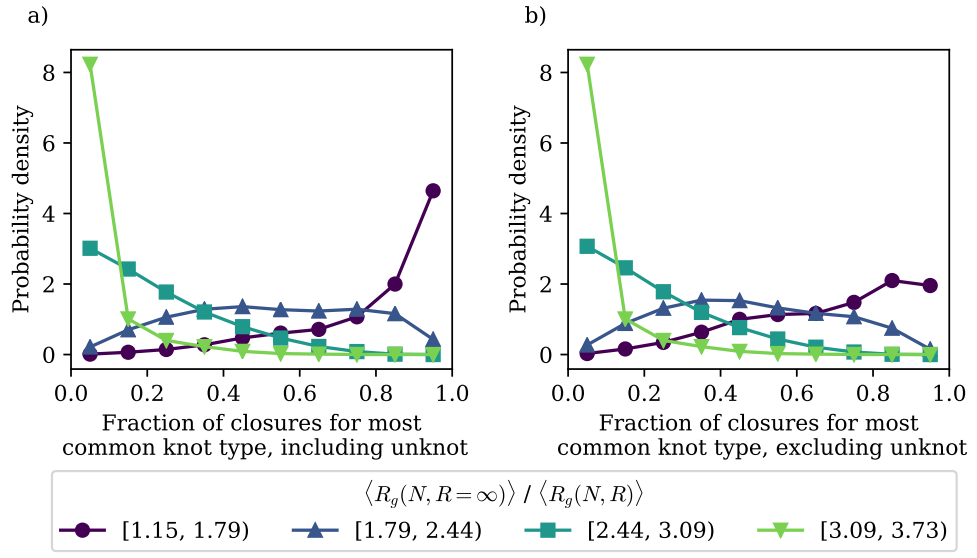


Figure 5.21: The distribution of most common knot coverage for walks in different ranges of degree of confinement. a) shows all walks while b) shows only walks whose most common knot is non-trivial.

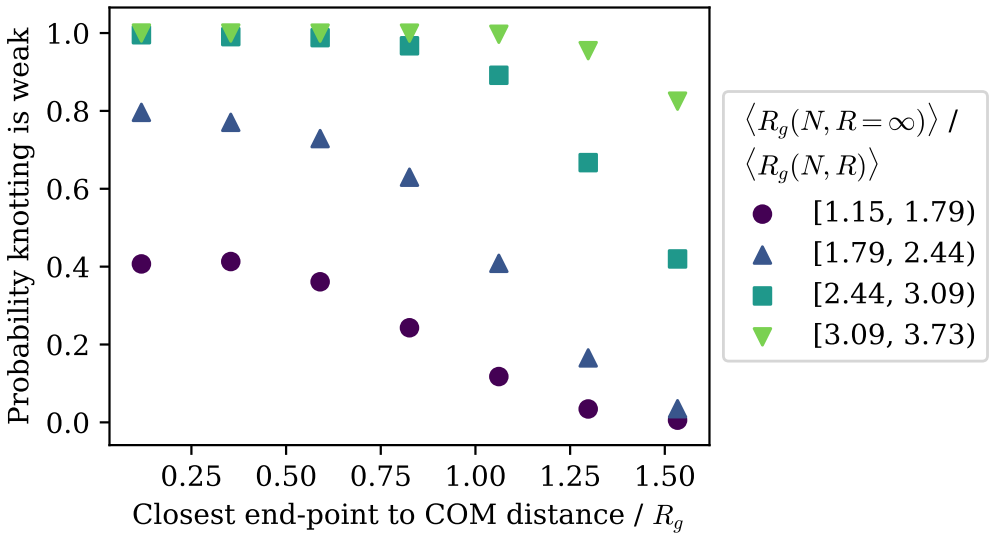


Figure 5.22: How the closest end-point to centre of mass distance over R_g affects the probability of knotting to be weak. Walks are binned by confinement degree, with ranges as indicated in the legend. Error bars of the order of marker size.

and the more complex that space curve is, the greater the variety of knots we expect to see across closures and hence the more weak knotting we expect. The greater the confinement degree, the more complex the walk is likely to be. This would explain the increased probability of weak knotting and the lower average coverage of most common knot seen at greater confinement degrees.

To investigate this more directly, we group the walks we have investigated by degree of confinement, and ask how the end-point position relates to weak knotting probability. For each walk we determine which end-point lies closest to the centre of mass of the walk, and divide this by the R_g of that walk. We bin the walks by this ratio and determine the fraction of the knotted walks in each bin which are weakly knotted. We plot this in Fig 5.22.

For all walks we see that the closer an end-point is to the centre of mass as a ratio of R_g, in other words the more buried an end-point is, the more likely the knotting is to be weak. We also see that the knotting in walks with higher confinement degree is more likely to be weak, even at the same relative end-point position. Even though the end-points are embedded by the same amount, the higher the confinement degree the more dense the tangle the end is being buried in. So we see that our hypothesis for how weak knotting arises is supported by our data.

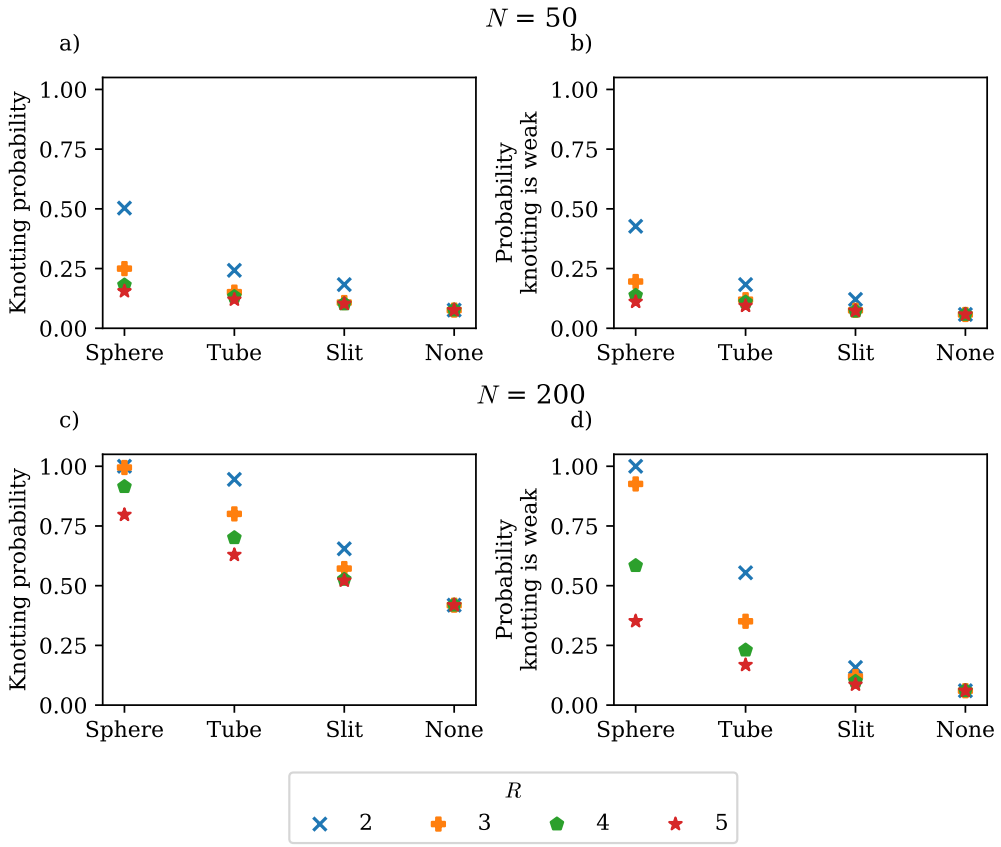


Figure 5.23: Knotting, a) and c), and fraction of knots which are weakly knotted, b) and d), for random walks in different confinement geometries. Plots a) and b) are for walks of length 50, and plots c) and d) are for walks of length 200.

5.2.3 Comparing off-lattice walks confined to spheres, tubes and slits

We now broaden our attentions to varying the geometry of the confinement. By comparing walks confined to spheres, tubes, slits and outside confinement, we vary the number of confined dimensions and investigate the impact this has on knotting.

We start by comparing the probability of knotting and the probability that the knotting is weak in each confinement. The results for the longest and shortest lengths investigated are shown in Fig 5.23. Radius here has its usual meaning in spheres and cylinders, and refers to half the width between the slits. In this way we keep the distance between opposite points of the confinement constant, and open up one dimension at a time.

We see the results for knotting probability in Fig 5.23 a) and c). Unsurprisingly, the walks confined to spheres are more often knotted than those in tubes of equivalent radius, which are more knotted than those in slits, which are more knotted than unconfined walks. This difference is more stark in the longer walks, up to the point that knotting saturates in the spherically confined walks. We also see that the same reduction of radius results in proportionally more knots in spheres than in tubes, and in tubes than in slits. Again, there is a length dependence in the baseline of knotting since longer unconfined walks are more often knotted than shorter ones.

The probability for knotted walks to be weakly knotted is shown in Fig 5.23 b) and d). Apparent are similar trends to overall knotting probability with spherically confined walks being the most knotted and showing the biggest response to decreased radius, followed by walks in tubes and then slits. Here, as before, the baseline for weak knotting does not change with length. It is quite stark however just how little difference length makes to the probability of knotting being weak for walks in slits. The walks confined to tubes show a much increased weak knotting probability, especially at the smallest radius of confinement. This effect is all but gone when moving to slits, at least at the radii investigated. We do not investigate small enough radii to observe a decrease in knotting as radius decreases [102, 105].

As before, we can plot this data against degree of confinement, as defined previously in Eq 5.14. This can be seen in Fig 5.24, where we have included all lengths and confinement radii investigated. Immediately obvious is that the range of confinement degree explored is much smaller for walks in tubes, and smaller still for walks in slits, than what we explored in spheres. Remembering that we used the same radii and lengths in each geometry, we see that decreasing confinement radius increases the degree of confinement much less the fewer dimensions are confined. We see a length dependent spread of results in knotting probability as before which is not present in the probability of knotting being weak.

However, what is most stark is the rise of knotting and weak knotting with degree of confinement is steeper for walks in tubes than spheres, and steeper still for walks in slits. For the same relative reduction in ensemble average R_g, we see more knots, and more of those knots are weak, the fewer dimensions are confined.

As before, we can ask not just how many of the knotted walks are weakly

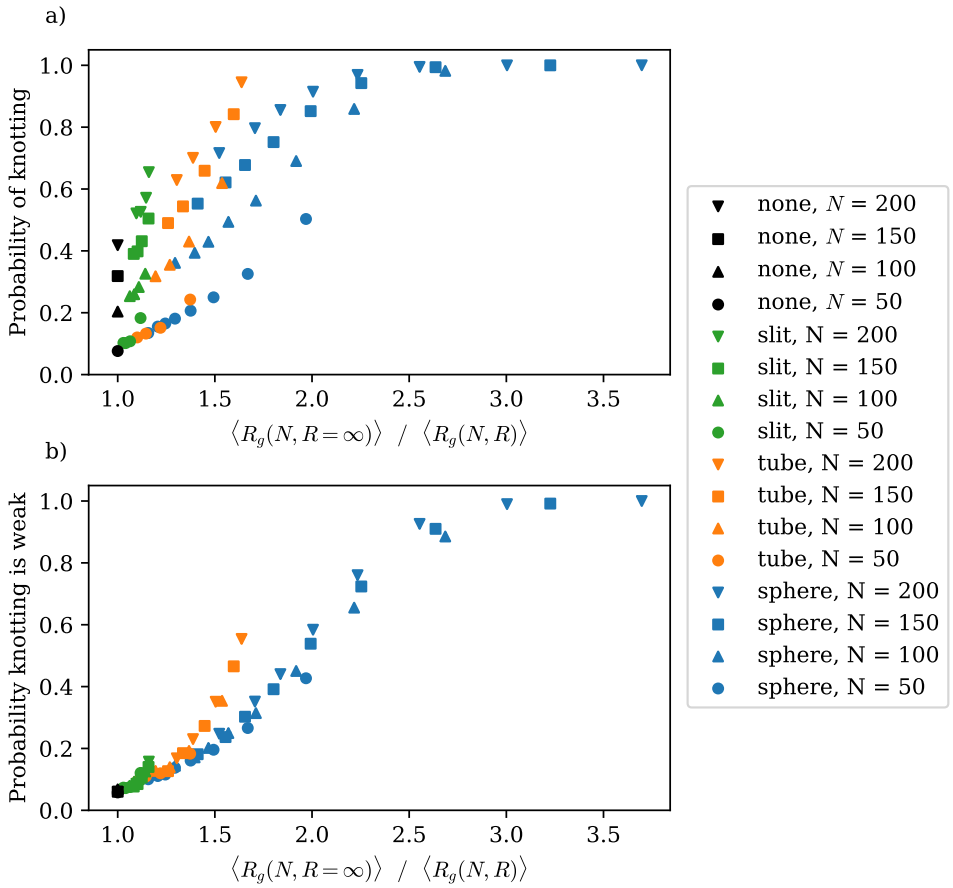


Figure 5.24: Probability of knotting, a), and probability that knotting is weak, b), with degree of confinement for the different confinement geometries investigated. None here means unconfined.

knotted, but how weakly knotted are the walks, characterised by the average coverage of the most common knot across closures. This is plotted in Fig 5.25. This data shows a similar trend as Fig 5.24, in that for the same change in confinement degree, there is a greater change in average coverage for walks in tubes than walks in spheres, and even more so for walks in slits. For the same confinement degree, walks in slits have the lowest average coverage of the most common knots, followed by tubes and then spheres.

The difference in behaviour with degree of confinement between the different confining geometries is attributable to the way we calculate the degree of confinement in the first place. We introduced the degree of confinement to gain a measure of how compact an ensemble of confined walks is, compared to how they would be outside of that confinement. We compare the radius of

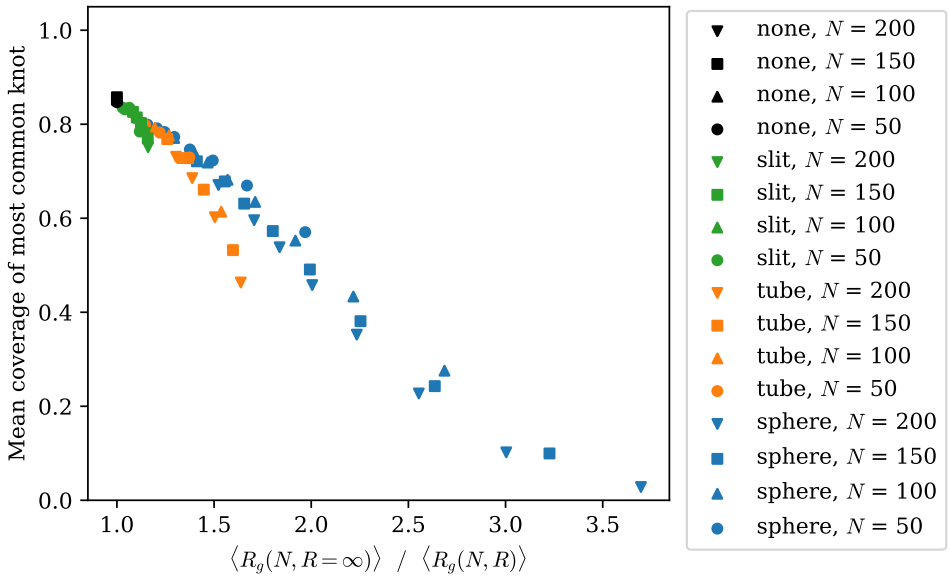


Figure 5.25: The ensemble average coverage of the most common knot in off-lattice walks in different confined geometries, plotted against degree of confinement.

gyration of these walks, which is a one-dimensional, scalar quantity. This works well for walks confined to a sphere, where the confinement is equivalent in all directions. When we move to tubes this is of course no longer the case. Walks in the tube are not likely to be compact in the ball-like way that walks in the sphere are. Their constraint is direction specific. This difference is highlighted even more when moving to the slit.

A more appropriate measure of the shape and size of the walks may be the *gyration tensor*, S as used by Rawdon et al. [150]. This is defined as:

$$S_{ij} = \frac{1}{2N^2} \sum_{n=1}^{N} \sum_{m=1}^{N} (r_n^i - r_m^i)(r_n^j - r_m^j) \qquad (i = 1,2,3;\ j = 1,2,3) \tag{5.15}$$

where N is the walk vertices, and r_n^i is the i^{th} coordinate of the n^{th} vertex. For example

$$S_{xy} = \frac{1}{2N^2} \sum_{n=1}^{N} \sum_{m=1}^{N} (x_n - x_m)(y_n - y_m) \tag{5.16}$$

We can find a coordinate system to diagonalise the resulting matrix, giving us:

$$\boldsymbol{S} = \begin{bmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{bmatrix} \tag{5.17}$$

where $\lambda_1 \leq \lambda_2 \leq \lambda_3$. The sum of these eigenvalues gives us the square of the radius of gyration, i.e.

$$R_g^2 = \lambda_1 + \lambda_2 + \lambda_3 \tag{5.18}$$

and the square roots of the eigenvalues give us the semi-axis lengths of the associated *ellipsoid of inertia* and hence tell us about the shape of the walk.

Rawdon et al. [150] choose a specific normalisation of the eigenvalues, where instead of dealing with λ_i, they use:

$$\alpha_i = \sqrt{3\lambda_i} \tag{5.19}$$

Normalised in this way, α_1, α_2 and α_3 give the semi-axis lengths of what the authors refer to as the *characteristic inertial ellipsoid*. In the same way as the radius of gyration of a sphere is its radius, the α_i's of an ellipsoid are its semi-axis lengths. For sake of easy comparison, we choose also to proceed using α_i's, rather than λ_i's.

Two key quantities related to shape can be defined from these semi-axis lengths. The *asphericity* of a walk tells us how equal the semi-axis lengths are [151], and the *prolateness* tells us whether the middle semi-axis is closer in length to the longest or the shortest semi-axis [152]. The asphericity is defined as:

$$A(\alpha_1, \alpha_2, \alpha_3) = \frac{(\alpha_1 - \alpha_2)^2 + (\alpha_1 - \alpha_3)^2 + (\alpha_2 - \alpha_3)^2}{2(\alpha_1 + \alpha_2 + \alpha_3)^2} \tag{5.20}$$

An asphericity of 0 means that the characteristic inertial ellipsoid is spherical, i.e. $\alpha_1 = \alpha_2 = \alpha_3$. The other extreme of asphericity is when $\alpha_1 = \alpha_2 = 0$, where the walk is a straight line and $A = 1$. Prolateness is defined as:

$$P(\alpha_1, \alpha_2, \alpha_3) = \frac{(2\alpha_1 - \alpha_2 - \alpha_3)(2\alpha_2 - \alpha_1 - \alpha_3)(2\alpha_3 - \alpha_1 - \alpha_3)}{2(\alpha_1^2 + \alpha_2^2 + \alpha_3^2 - \alpha_1\alpha_2 - \alpha_1\alpha_3 - \alpha_2\alpha_3)^{3/2}} \tag{5.21}$$

which ranges between -1 for a perfectly oblate ellipsoid where $\alpha_2 = \alpha_3 > \alpha_1$ and 1 for a perfectly prolate ellipsoid where $\alpha_1 = \alpha_2 < \alpha_3$. Between asphericity, prolateness and radius of gyration we have good descriptors for the overall size and shape of the walks. Using these quantities, it was confirmed by Rawdon et al. [150] that the average shape of unconfined random walks is aspherical in a prolate way, but that the shape of knotted walks was less aspherical, or more spherical.

The variation of R_g, A and P in confined walks as the confining radius is reduced is shown in Fig 5.26. The top row, a), shows how the radius of gyration varies. Walks in all confinement shapes show a reduction in R_g, with the greatest

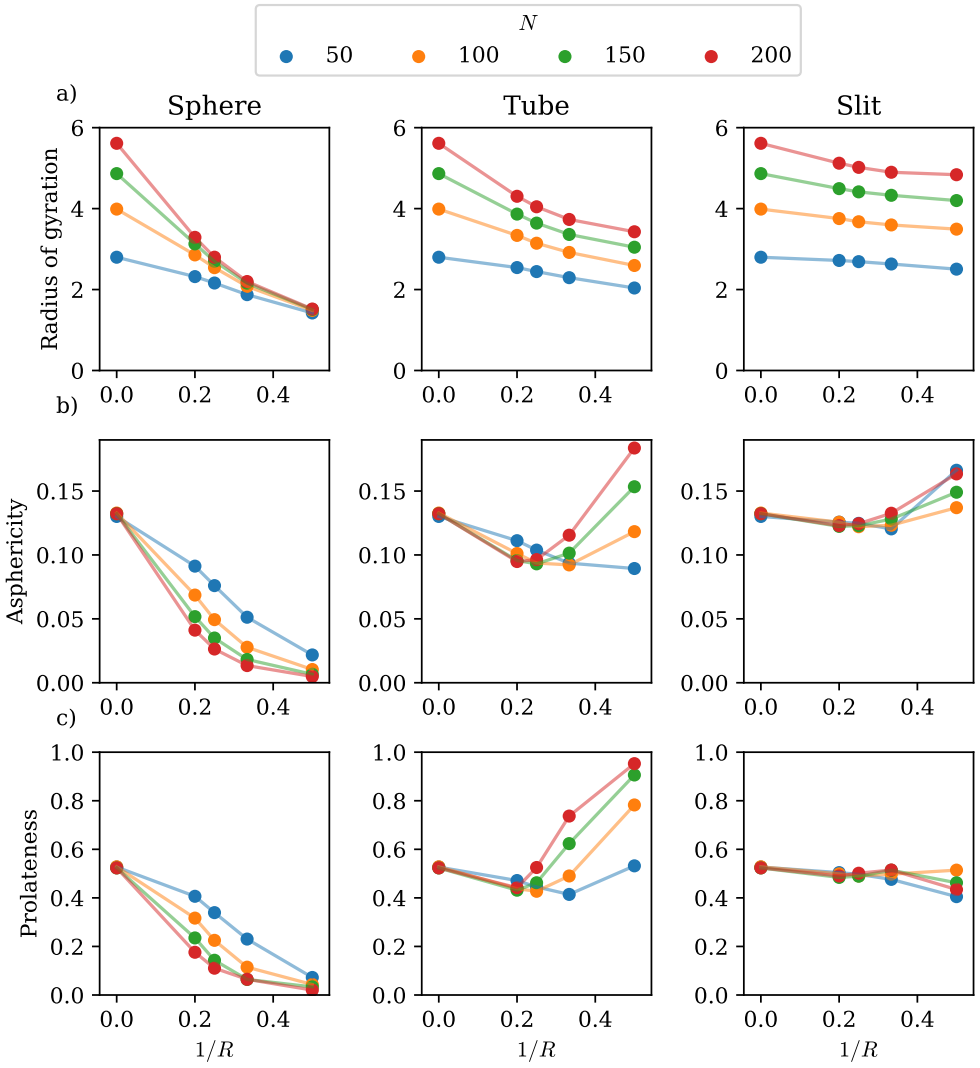


Figure 5.26: The mean a) radius of gyration, b) asphericity and c) prolateness of random walks confined to spheres, tubes and slits with inverse confining radius, R.

reduction for walks in spheres where walks of all lengths converge to the same R_g as confinement tightens. Clearly the R_g of these walks is limited by the sphere. In the case of walks in tubes and slits, there is no hard limit on R_g and so there is more length dependence, even in the tightest confinements. We see the R_g of tubes is restricted more than for walks in slits as we know must be the case from Fig 5.24.

What of the shape of the walks? The middle row, b), of Fig 5.26 shows the asphericity of the walks against $1/R$. For walks in spheres, the asphericity

tends to zero (a spherical shape) as the confinement tightens, with the fastest approach for the longest walks. Walks in tubes and slits are more interesting. For the tightest confinements we would expect the walks to be more aspherical than unconfined walks as the walks are forced to take on the shape of the confining volume. We see that this is likely to be the trend, but at the confining radii investigated we first see a slight reduction in asphericity. This is likely because the confinement is too large to strongly affect walks of this length, and so the walks are not perfectly aligned with the confining volumes. This first results in a squashing of the unaligned walks, and so a reduction in asphericity at large confining radius and short lengths, after which the alignments agree and the walks take on more of the shape of the confinement. This effect is more dramatic in tubes, whereas in slits the change to asphericity is very minimal.

Finally we plot prolateness in the bottom row, c). We expect that walks in spheres will tend to a prolateness of zero, walks in tubes to a prolateness of one, or prolate, and walks in slits to be a little more oblate than unconfined walks. This is essentially what we see, although again there is an interesting non-monotonic behaviour in tubes, which initially fall in prolateness before rising again. There is very little variation for walks in tubes.

We showed earlier knotting and weak knotting in all confinements increases monotonically as confining radius decreases and length increases, although this is not the case indefinitely [102, 105]. As prolateness and asphericity both show non-monotonic behaviour, they alone cannot predict the increase in knotting and weak knotting we see with longer lengths and tighter confinements. We must take account of the relative sizes of the walks as well as their shapes.

To do this, let's look at how the semi-axis lengths vary with walk length, confining radius and confining shape. The mean values of the semi-axis lengths are shown in Fig 5.27 for walks of 50 and 200 steps. The top row of graphs, a), shows the absolute value of the semi-axis lengths, and the difference between the confinements is clear and unsurprising. In spheres, all three semi-axes are reduced as the confining radius tightens; in tubes, while all three are reduced to an extent, the largest semi-axis, α_3 is not reduced nearly as much as in the sphere; and in slits, the reduction in size is minimal for all three semi-axes, at least at the range of confining radius explored here. The second row of graphs, b), shows how the semi-axis lengths vary relative to the unconfined lengths. In spheres, α_3 is the most reduced relatively, followed by α_2 and then α_1; short walks in tubes show a similar trend to a lesser extent, but interestingly in the

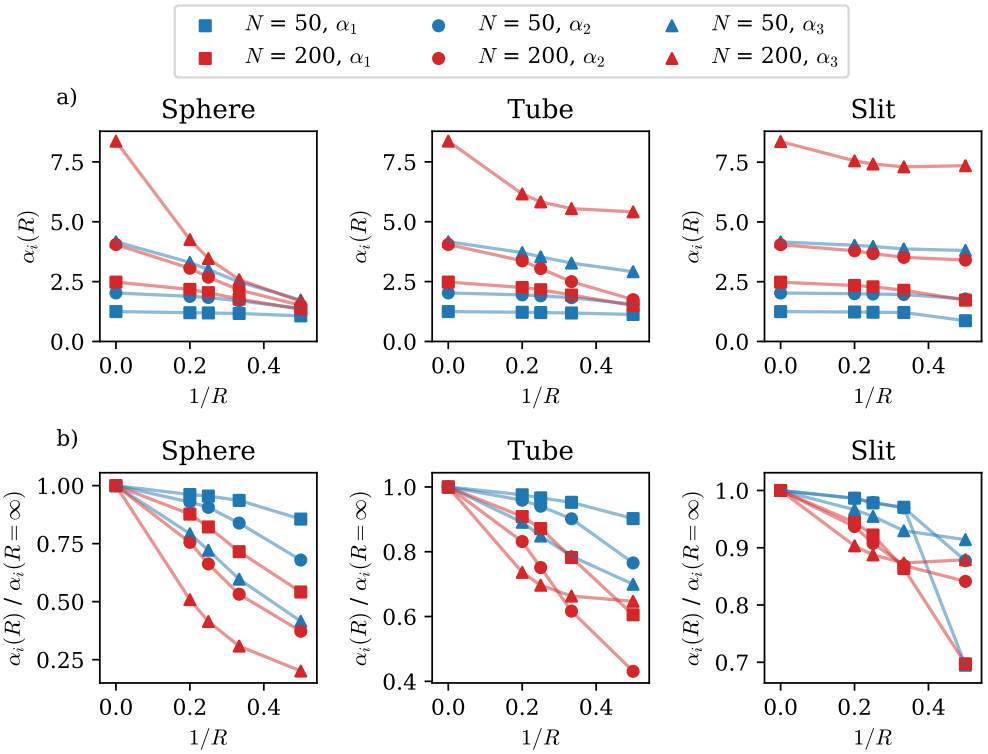


Figure 5.27: The mean semi-axis lengths of the characteristic inertial ellipsoid of random walks confined to spheres, tubes and slits. The first row, a), shows the absolute values and the bottom row, b), shows the values relative to the unconfined lengths.

longest walks, α_3 is the least reduced relatively despite being the most reduced in absolute terms; and in slits we find this inversion even for the shortest walks.

With degree of confinement, we compared the radius of gyration of walks. The radius of gyration is formed from a sum of semi-axis lengths. Considering the longest semi-axis is significantly longer than the others in all but spherically confined walks, the degree of confinement is dominated by its contribution. As we've just seen though, α_3 in tubes and slits shows the least reduction of all the semi-axes compared to unconfined walks. A different measure which places each semi-axis on the same footing is the following *adjusted degree of confinement*:

$$\text{adjusted degree of confinement} = \frac{1}{3}\sum_{i=1}^{3}\frac{\langle\alpha_i(N,R=\infty)\rangle}{\langle\alpha_i(N,R)\rangle} \tag{5.22}$$

where the mean across confined walks of each semi-axis is compared separately to that of the unconfined walks before they are summed. We divide by three

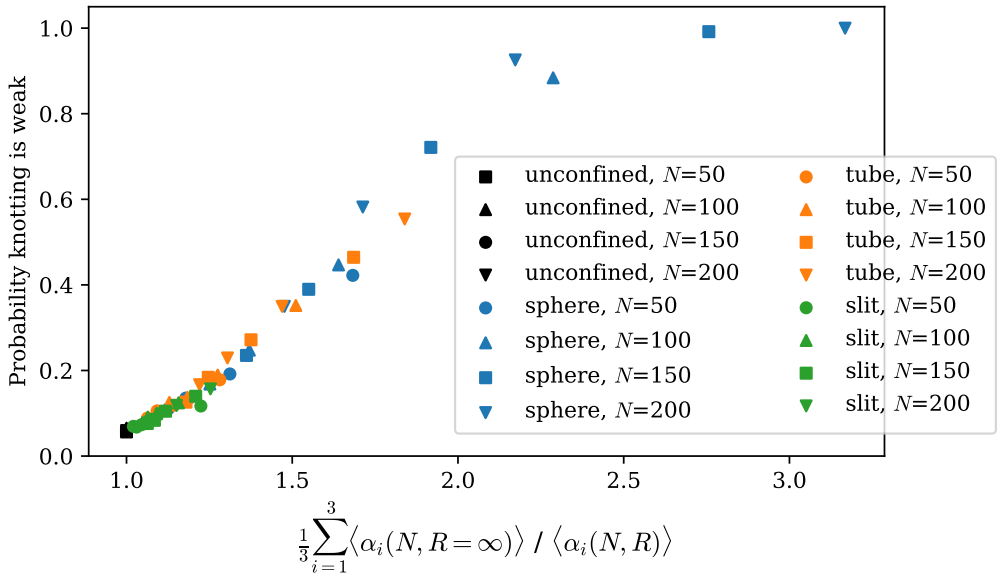


Figure 5.28: Probability that knotting is weak against the adjusted degree of confinement for walks confined in spheres, tubes and slits.

just so that the adjusted degree of confinement for unconfined walks is one.

When we plot the probability that knotting is weak against the adjusted degree of confinement, we get Fig 5.28. As can be seen, all our data points now lie on the same line, within a certain degree of scatter. The error here is small and so any deviations must be due to length or other factors we do not account for here. We can also plot the mean coverage of the most common knot against the adjusted degree of confinement, Fig 5.29, again seeing that all the values fall onto the same line. When we plotted these quantities against degree of confinement earlier we saw that in tubes and slits, there was a larger change with a small change in degree of confinement than compared to spheres. As the degree of confinement was dominated by the α_3 contribution, and the reduction of α_3 was relatively small, the effects of the relative reductions in α_1 and α_2 were obscured. Since the adjusted degree of confinement puts all the α's on an equal level, we see that the relative reduction of all the semi-axes plays an important role in the weak knotting of confined walks.

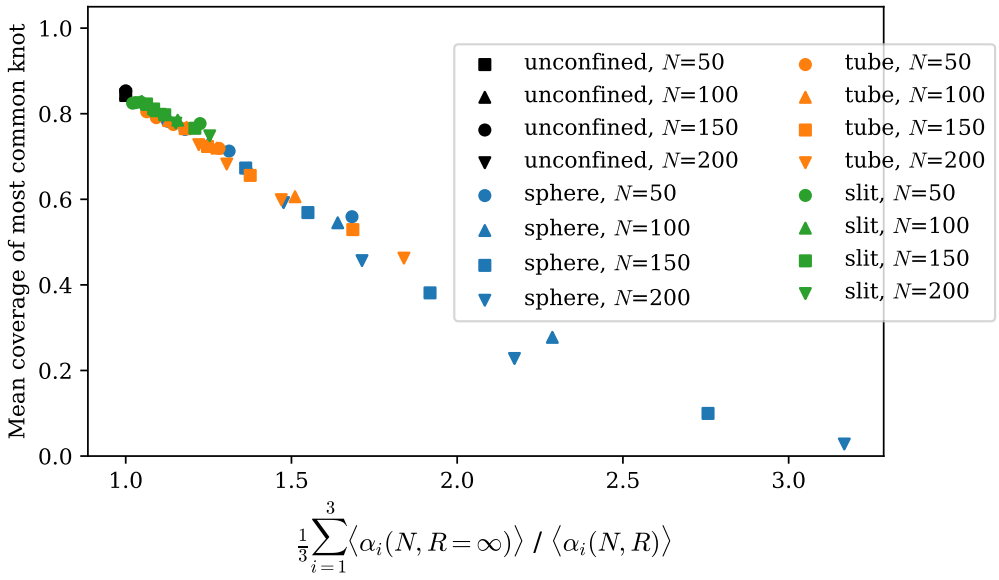


Figure 5.29: Mean coverage of most common knot against the adjusted degree of confinement for walks confined in spheres, tubes and slits.

5.2.4 Robustness of knotting results in off-lattice walks confined to the tube

In Sec 5.1.2 we covered how we generate random walks confined to tubes. We had to choose a value of n which determined how much the boundary avoiding effect of an accept-reject walk was compensated for. The results presented so far have been for a value of $n = 0.6$, as this provided a good balance of not having the vertex density in any particular part of the tube too different from uniform, while keeping the overall difference from uniform low also. We have duplicated all of these results for walks with a value of $n = 0.5$, which produced a vertex density with both larger maximum and total differences from uniform. Here we present how this affects the knotting of the walks.

The behaviour of knotting and weak knotting with degree of confinement is given in Fig 5.30. The overall behaviour here has been discussed previously. The important feature to highlight here is how close the walks with $n = 0.5$ and $n = 0.6$ lie. In all cases, the markers for comparable parameters overlap considerably. As throughout this section, the error bars are comparable in size to the markers and so much of the difference can be attributed just to error.

Further to this, we can examine the average coverage of the most common knot over closures in Fig 5.31. Again here we see that the values for walks with

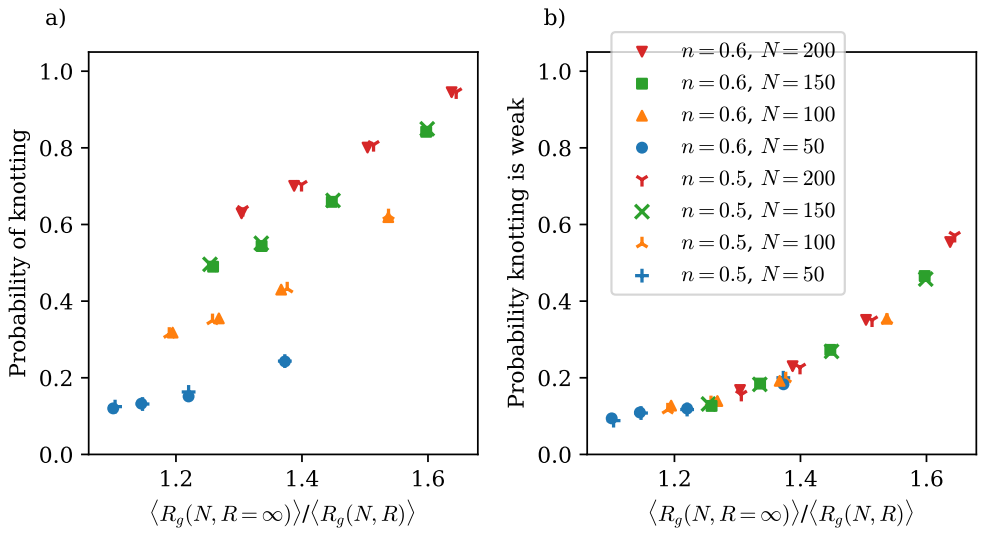


Figure 5.30: The probability of knotting, a), and probability that knotting is weak, b), for random walks confined in the tube, with a boundary behaviour value of $n = 0.5$ and $n = 0.6$.

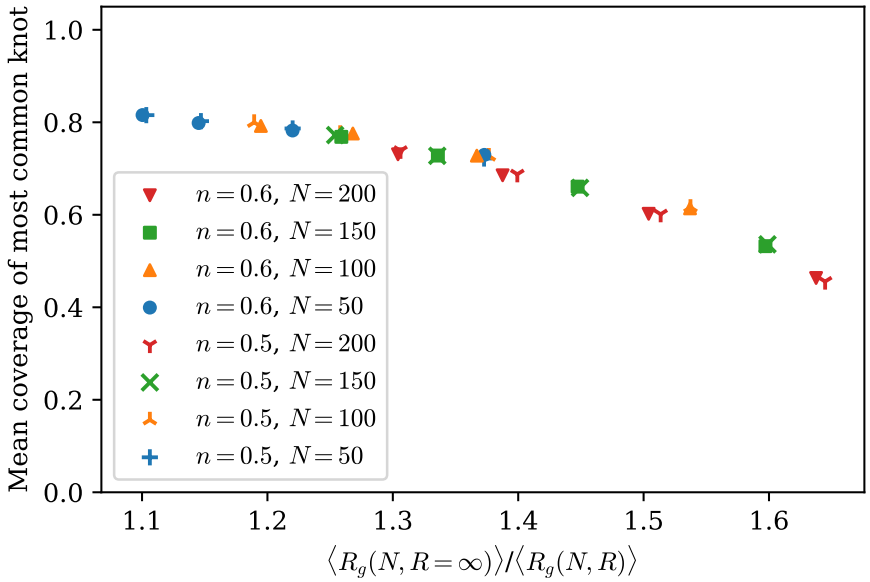


Figure 5.31: The ensemble average most common knot coverage over closures for random walks confined in the tube, with a boundary behaviour value of $n = 0.5$ and $n = 0.6$.

$n = 0.5$ and $n = 0.6$ are very close to each other, both displaying the tendency for the most common knot to cover fewer closures as degree of confinement increases.

While there are differences here between the two walk types, they are very small and almost within the margin for error. From this we conclude that the knotting statistics for a truly uniform random walk in a tube are likely to be very close to those presented in this chapter.

6 Conclusions and discussion

6.1 Method and results summary

In this thesis we have addressed the problem of how to recognise and classify knotting in open curves. We took inspiration from the procedures in [20, 26, 27, 39, 118, 23] which involved joining the ends of open curves with straight lines to uniformly distributed points on a large surrounding sphere, a method we called sphere closure. Our approach was to instead take projections of the curve from uniformly distributed directions and join the ends in the projected diagrams with virtual crossings. This procedure, which we called virtual closure, could result in a classical knot or a virtual knot on projection. We conceptualised the virtual knots as lying 'in-between' classical knot types, allowing finer distinctions to be made between the knotting of different open curves. We discussed how we distinguish different virtual knot types in a computationally efficient manner using the generalised Alexander polynomial and Jones polynomial.

Having introduced the method of virtual closure, we discussed the possible results one might expect for a given open curve. We divided the results broadly into: strong knotting, where a single non-trivial knot type appears in 50% or more closures; weak knotting, where no single knot type dominates, but the unknot still appears in 50% or fewer closures; and unknotting, where the unknot is formed in over 50% of closures. Different flavours of strong and weak knotting were highlighted, but the main distinction remained between those curves which strongly resemble a particular knot type, and those whose knot type is ambiguous, although they are still appreciably tangled. Given that there are more possible knot types open to virtual closure than to sphere closure, we expected weak knotting to be more likely in a virtual closure analysis.

With virtual closure in hand, we moved to our analysis of knotting in proteins

deposited in the PDB. We followed the cues of previous protein knot surveys [23] in selecting the largest unique set of chains to analyse, totalling around 160,000 chains. Reproducing their results under sphere closure, we found approximately 950 protein chains to be knotted. Using virtual closure, we increased the number of chains defined to be knotted to around 1,250, an increase of almost 25%. The vast majority of the newly classified knotted chains displayed weak knotting, with most closures giving either the classical trefoil, 3_1 or the virtual trefoil, $v2_1$. There was also a reclassification of around 230 chains strongly knotted under sphere closure as weakly knotted under virtual closure, showing the knot type of these chains was perhaps not as certain as was previously thought.

In order to understand how typical this weak knotting in proteins may be and to investigate how weak knotting arises generally, we analysed the knotting of confined random walks. Two models of random walk were used: a confined, self-avoiding cubic lattice walk generated from Hamiltonian walks [22], and an off-lattice ideal chain walk. We discovered the hard way that if the off-lattice walks are confined using an absorbing boundary condition, they will be less likely to lie within a unit of the boundary than uniformly distributed points. To compensate for this and reproduce a uniform distribution of walk vertices, we used the method given in [149] to confine the off-lattice walks to a sphere. We adapted this method for use in tubes and slits, noting that the distribution of points is not quite uniform for walks in tubes. The difference was not thought to significantly affect knotting statistics.

In comparing confined lattice and off-lattice walks, it was found that while overall knotting was quite different in each case, with off-lattice walks being significantly more likely to knot at a given length, the likelihood that the knotting was weak was very similar. This weak knotting likelihood rose with length, in stark contrast to the unconfined off-lattice walks which showed negligible weak knotting at all lengths. It seemed that prevalence of weak knotting is confinement dependent in the main and not model dependent.

Investigating spherically confined off-lattice walks in more detail, we saw that the probability that the walks are knotted depends both on length and the degree of confinement, while the probability that this knotting is weak depends almost entirely on the degree of confinement. We introduced the degree of confinement as a measure of how compact confined curves were in comparison to their unconfined equivalents, with more confined walks more likely to knot and for that knotting to be weak. It was also seen that the average coverage

of the most common knot, a measure of how well represented the knotting of the walks are by a single knot type, also depended almost entirely on degree of confinement. By analysing the position of the end-points relative to the bulk of the curve, we showed that the closer an end-point is to the centre of mass of the curve, and the more compact that curve is, the more likely the knotting is to be weak.

The picture became more complicated when we reduced the number of confined dimensions, moving to walks in tubes and slits. At equivalent confinement radius, the probability of knotting was less in tubes and even lower in slits, compared to spheres, with a similar trend found in weak knotting. Plotting these quantities against an adjusted degree of confinement which placed compression in each Cartesian dimension on equal footing brought all the results onto the same curve.

6.2 Results discussion

Now that we have presented the results, we can take a broader view over them and tie some parts together.

6.2.1 Proteins in context

One motivation behind the random walks work was to understand how the knotting seen in proteins compared to generic open curves. It has already been noted that proteins are less frequently knotted than geometrically similar compact random walks [146]. Our results from virtual closure are essentially the least surprising possibility, find nothing to dispute this fact. We found neither that many more proteins are knotted under virtual closure, or very few. Instead we see that there is a small but significant proportion of protein chains which are somewhere in-between being a clear unknot and a clear trefoil knot, a tangling which sphere closure was not sensitive to but that virtual closure was. Still, fewer than 1% of protein chains were found to be knotted.

However, as has been highlighted often in this thesis, one of the biggest differences between sphere and virtual closure is the sensitivity of virtual closure to weak knotting, or open curves with significant ambiguity in knot type. With many of the already known knotted protein chains being reclassified as weakly knotted under virtual closure, we found almost 40% of knotted chains to be

weakly knotted. There is no equivalent in the random walks results of a walk so infrequently knotted, yet with so many of those knots being weak.

This is a surprising fact. We have frequently associated weak knotting to the end positions of open curves. We showed that having end-points within the bulk of the curve increases the probability that curves are weakly knotted. However, we expect the ends of proteins to lie close to the surface of the proteins [21], leading to strong knots. Of course, it would be highly unusual for the ends of a protein to lie very far outside the rest of the protein, so the knotting of individual chains is unlikely to be as strong in the extreme as certain random walk configurations. When plotting the end position to centre of mass distance of the knotted proteins there was not a notable excess with ends very close to the centre though. Considering how most weak knots in proteins have a significant unknotted component, and the weak knotting is only brought out under virtual closure, it is possible that most of the weak knotting we see is due to the partial tucking of an end under part of the backbone chain as was highlighted by Mansfield when he first looked for knotted proteins [20].

If one insists on trying to find the closest equivalent random walks for the weak knotting in proteins, we can leverage the fact that the length distribution for these chains is sharply peaked around 265 amino acids and try to find the best fit random walk length. We chose to use the lattice walks we did due to their previous comparison to proteins [146]. Looking at these walks then, the weak knotting of proteins is best matched by walks on a $6\times6\times6$ lattice of length around 110 steps, although these show 4 or 5 times more knotting overall than proteins.

We could instead assume that the proteins are relatively spherical, rather than tubular or planar, and attempt to find an equivalent degree of confinement. We only have data for off-lattice walks here and so the comparison is not quite as valid as the lattice walks, but a degree of confinement comparison may be more meaningful than a single equivalent length. Spherically confined off-lattice walks with a degree of confinement around 1.8 give the best fit for proteins from our data. However, the underlying distribution of most common knot coverage is much more sharply peaked for proteins than for random walks as shown in Fig 6.1. The degree of confinement of proteins could be accessed more directly by comparing disordered proteins in theta conditions to equivalent length, ordered proteins.

An important distinction in protein knots is between deep and shallow knots.

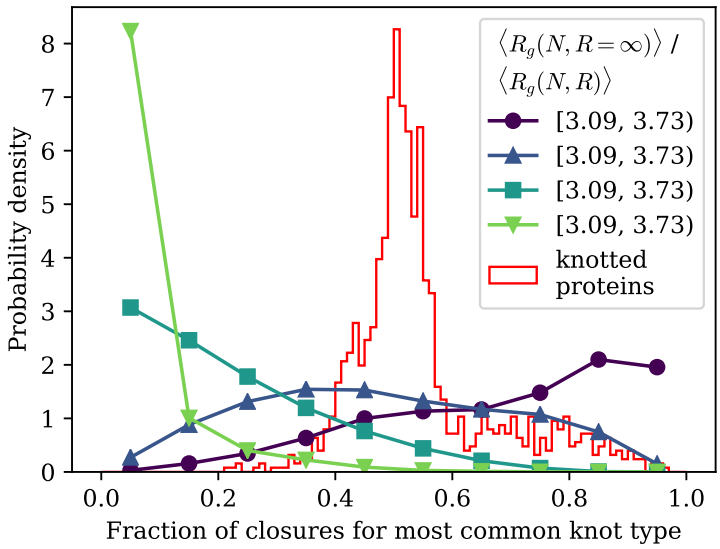


Figure 6.1: The distribution of most common knot coverage in spherically confined random walks compared to knotted proteins. Only walks whose most common knot is non-trivial are included for the random walks.

The discovery of deeply knotted proteins marked a step change in researchers taking knotted proteins seriously. While it is not a direct measurement of knot depth, shallow knots and weak knots are closely related. It perhaps would not have surprised earlier researchers to find that, while proteins do indeed knot, their knotting is more frequently weak than similarly knotted random walks.

Of course, we should be wary of treating proteins in general and proteins in the PDB in particular as some sort of representative ensemble. While the evolutionary exploration of protein space may sample from a similar space of conformations to random walks, ordered protein structure is not random. If a particular knot exists in a protein it exists for a purpose, and there is no reason to believe that the most common sorts of knots in random walks are the most useful for biological function aside from, perhaps, the ease of formation. Additionally, the PDB contains many structures of the same proteins, either in different environments or with slight mutations. The frequency of each knot is highly influenced by this and again is far from random. Given this, the value of a straight statistical comparison like this is limited and a more detailed study is needed.

6.2.2 Reflections on virtual closure

Given that virtual closure is a more abstract method for knot recognition than classical closure, taking longer and producing results which are somewhat harder to interpret, a less invested researcher may ask why use virtual closure at all? Indeed, there are many examples of open curves where virtual closure and sphere closure give very similar results. Typically these will have a well defined, locally contained knot, with long loose ends either side. The typical shoelace or necktie knots are good examples, but also many unconfined polymers where knots are known to localise. Of course, other methods such as radial closure or minimally interfering closure would also give the same results here.

In general, the advantage of the statistical closure methods where many different closures are considered is that consistent answers can be obtained for less clear-cut configurations. These answers will be robust to small, topologically inconsequential perturbations that single closure methods may not be. Typically, these difficult conformations will have end-points which are close to or within the bulk of the curve, where a single closure which most captures the essence of the tangling is difficult to find. The first question one should ask when choosing a knot recognition method then is, are my ends likely to be loose and free from where I expect the knot to be, or not. This may be easy to answer if there are few curves to analyse and where this evaluation can be done by eye, or if the end-points are anchored at a barrier through which the curve cannot pass.

If the end-points are not likely to be so free of the tangling then a statistical closure method should be used. It is useful to think about what each method actually measures. In sphere closure, each closure represents the knot created if the ends were extended directly to meet at a point outside the curve. For a flexible enough system, this corresponds to pulling the ends out of the curve and joining them. This can have physical relevance, for instance if one is trying to tie a particular knot, say with optical tweezers or by biasing a folding pathway. Sphere closure can tell the direction needed to pull the ends to produce this knot.

Virtual closure on the other hand, tells what the curve looks like currently from any given direction, without any more physical manipulation of the curve or addition of information. In order to do this, we have to extend the knot types we use to include virtual knots. The most intuitive way we've found to think

about virtual knots in this context is as in-between classical knots. A projection which is not quite a trefoil and not quite a cinquefoil will have this quality captured by a single virtual knot. This adds a sensitivity to virtual closure over sphere closure, allowing it to distinguish more finely the types of tangles seen.

Much of the analysis of statistical closure data previously has been based on finding a dominant knot type across closures. Of course, it hasn't escaped notice that this is not always applicable [26, 28], but usually the curves analysed have presented a dominant knot. With the introduction of virtual closure and the additional knot types available, single representative knot types are less common in open curves, particularly in the less clear cut conformations we recommend a statistical closure method for. From our results, we see that virtual closure discriminates more finely between curves with ambiguous knot type. Weak knots are declared for less complex curves, and the degree of ambiguity, gauged by the coverage of the most common knot, rises faster for the same increase in curve complexity for virtual closure than sphere closure. An interesting consequence of this is that the weakness of knotting essentially saturates at a certain curve complexity for virtual closure, sooner than for sphere closure. This means that sphere closure may be a better discriminator of ambiguous knotting for extremely complex curves using the same number of closures. Of course one can take arbitrarily many closures to recover this power in virtual closure.

To answer the question when should virtual closure be used, we have a few recommendations. If the knot is highly localised and the end-points extended from the bulk, it likely does not matter which knot recognition method is used. Virtual closure will give the same answer as any other method but with an increased computational cost over single closure methods. In more complex curves where the end-points participate more in the tangling of the curve, virtual closure will allow a greater distinction between types of entanglements. Given its increased sensitivity to ambiguity, it will not confidently report a single dominant knot type unless that is clearly the case. If this nuance is not needed, or if the knot type obtained when extending the ends in a particular direction is required, sphere closure will suffice. It should be noted however, that if sphere closure is performed using the Jones polynomial there will essentially be no computational benefit over virtual closure, as the Jones is also an invariant of virtual knots.

A background concern in much of the results is the arbitrary nature of using 50% coverage as a cut-off in knot classifications. If there has to be a cut off, 50%

is as natural as any, but clearly there is a big difference between a curve with a trefoil in 90% of closures, and one with a trefoil in 60% of closures. Similarly, there is little conformational difference between curves showing 51% and 49% trefoil coverage.

Depending on the type of analysis being done, there are more nuanced ways of tackling the ambiguity of knot type. If very few curves are being analysed, dealing with the complete knot spectrum of each curve will provide great detail. When looking at a complete spectrum, the researcher can decide for themselves what the significant knot types are and which components are of interest. If an ensemble of curves is being analysed, the distribution of most common knot coverage is useful. This provides more information than the binary classifications of strong and weak, but in a manageable format. When analysing proteins in this way, we found the distribution of most common knot coverage peaked just above 50%. A slight perturbation of protein structure could have given very different results. The random walks showed the diversity possible in these curves, with more compact and complex curves having a distribution weighted towards no single knot covering many closures.

6.3 Knotoids

A significant development since the publishing of our first paper [113] has been the introduction of *knotoids* as a tool for recognising knotting in open curves. Knotoids, first introduced by Turaev in 2010 as a preprint and later published in 2012 [153], are essentially open knot diagrams. They allow the distinguishing of different open knot diagrams without closure. Several examples are shown in Fig 6.2. These diagrams can be manipulated by planar isotopy and the classical Reidemeister moves like ordinary knot diagrams, provided the open ends do not pass a strand and add or remove a crossing, much like in Fig 3.1 b). This allows knotoid diagrams to be topologically distinct from each other.

Knotoids are divided into *knot-type* knotoids, which have their end-points in the same face of the diagram, and *pure* or *proper* knotoids which do not. In Fig 6.2, a), d) and e) are knot-type knotoids and b) and c) are proper knotoids.

The connection to open curve analysis is relatively obvious from here. Projections of the curve can be taken, as is done in virtual closure, but instead of closing the ends with virtual crossings, the diagram is understood directly as a knotoid. All the same considerations about dominant knots and ambiguity

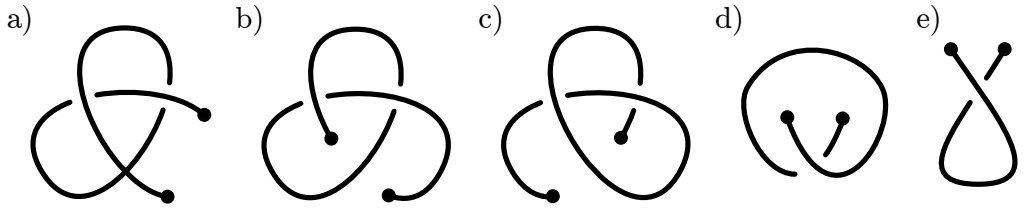


Figure 6.2: Knotoid diagrams. a) a knot type knotoid. b) and c) proper knotoids. d) a trivial spherical knotoid, but a non-trivial planar knotoid. e) a trivial knotoid.

of knot type that we have discussed are present here also. This was done by Goundaroulis and others in [124], where they performed a knotoid analysis of a number of proteins.

Virtual knots and knotoids are very closely related and behave in similar ways. A virtual knot can be obtained from a knotoid by performing a virtual closure of the diagram. This does not change the Gauss code of the knotoid, but it does remove some information about the underlying knotoid. For example, the knotoids shown in Fig 6.2 b) and c) are distinct as knotoids. No sequence of Reidemeister moves can deform one into the other. However, as their Gauss code is the same, they virtually close to the same virtual knot. In this way, knotoids are more powerful than virtual knots in distinguishing projections of open curves.

There is a clear similarity between the knotoids b) and c), and the only ambiguity in the virtual knot resulting from their virtual closure is which strand involved in the virtual crossing was added between the end-points. As far as we know, there are no known examples of distinct knotoids which close to the same virtual knot in a manner more complex than this. Additionally, the invariants used in [124] could not make this distinction, and so the analysis, while philosophically distinct, was no more powerful than a virtual closure analysis.

The comparison to virtual knots we have just made assumed that knotoids exist in $\mathbb{S}^2$ just as virtual knots do. This need not be the case, and in fact a theory of knotoids in $\mathbb{R}^2$ has also been proposed [125]. A distinction can then be made between knotoids in $\mathbb{S}^2$, which are called *spherical knotoids*, and knotoids in $\mathbb{R}^2$, which are called *planar knotoids*. The main consequence of moving to planar knotoids is that many previously trivial knotoids are now non-trivial. For example, Fig 6.2 d) and e) are both trivial spherical knotoids, but d) becomes

a non-trivial planar knotoid as the upper loop cannot be passed over the endpoints, and can no longer be passed around the back of the sphere to arrive at the other side. Planar knotoids then are the most powerful topological objects so far for classifying open curves, providing many more distinct topological classes than either virtual knots or spherical knotoids.

The proposing authors have provided software to perform knotoid analyses [154]. A significant part of the work in providing this software was the construction of a table of distinct spherical and planar knotoids which until now has not been available.

A full analysis of the PDB has recently been undertaken yet using planar knotoids by KnotProt. There is no paper currently accompanying this and the KnotProt results include backbones which are now redundant (in Chapter 4 we simply ignored these and adjusted the KnotProt results accordingly). In addition, they only take the most common knotoid as representing the knotting of the chain and so the comparison to our results is difficult with the information presented. As of July 2018 they had found 1026 chains with a nontrivial planar knotoid as its most common projection. It would be interesting to see how the results would change if our strong and weak knotting categories were used instead.

With these new tools then, what is the place of virtual closure? The knotoid approach captures the essential spirit of using virtual knots in the first place by analysing purely the open curve as is, without the addition of more physical closures. A spherical knotoid analysis would provide very similar answers to a virtual closure analysis, and if the cost of calculating invariants that can distinguish situations like Fig 6.2 b) and c) is not great, there is essentially no reason not to use the spherical knotoid approach. We expect that any change in the results of this thesis would be marginal if spherical knotoids were used.

Planar knotoids do provide a considerable increase in distinguishing power. It is not clear currently how meaningful this is as far as providing a measure of entanglement is concerned. Clearly, Fig 6.2 a) is more entangled than b), and b) more than d). The difference between d) and e) seems intuitively less significant topologically. Certainly planar knotoids will provide a finer classification of curve conformations than spherical knotoids. It would likely be system dependent whether this distinction made a difference to the system properties.

Compared to our analysis we would expect to see curves more frequently classified as knotted using planar knotoids. We would also expect to see a higher

proportion of these knots classified as weakly knotted and for the most common knot(oid) coverage to reduce. This may tip the balance in the case of proteins to having a majority of proteins weakly knotted. As for the degree of confinement of random walks, we would anticipate that the growth of weak knotting fraction with confinement degree would be steeper and saturate sooner than under virtual closure.

It should be noted here that we didn't use virtual knots to their full potential in this thesis. As mentioned, we do not distinguish between any mirrors of virtual knots. Given that there are two distinct mirrorings of virtual knots, horizontal and virtual, and these can be combined to give a third mirror, there are four different forms of each virtual knot related by mirrors which we class as the same knot. Not all virtual knots are distinct from their all mirrors, but many are distinct from their chiral partners, i.e. horizontal mirrors. If we included all mirrors as distinct we would expect a similar shift in our results as just described for planar knotoids, but to a lesser extent.

As a final note, knotoids can contain virtual crossings. Knotoids with virtual crossings are called virtual knotoids, and those with only classical crossings are called classical knotoids. This is not a consideration when projecting open curves as we do here as only classical crossings will be produced.

6.4 Future work

The work in this thesis marks the first time virtual closure has been used and the first time the full spectrum of knot types in a closure analysis has been studied in such depth. Naturally, there remain many questions to answer and avenues to explore. A simple step to take before embarking on future research would be to implement the curve simplification scheme used in [124] for virtual closure. While computational time has never been a major bottleneck in this project, this would be a great quality of life improvement to have.

An extended minimally genus one virtual knot table would be another useful addition. Currently, we distinguish minimally genus one virtual knots from others by eye. We only go up to four crossings as this is what Green's table provides diagrams for [112], and we very quickly leave this table as curves become more complex. Invariants are provided for higher crossing number but this is not useful alone.

There is a paper on genus one virtual knots [114] which provides a possible

solution to this. As well as going up to five crossings, they introduce a generalised Kauffman bracket polynomial which is in terms of two variables, a and x. From the polynomials given in [114], minimally genus one virtual knots always have a factor of x in front of every term, and no higher powers of x. Knots which are not minimally genus one have both higher and lower powers of x present. It is not proven that this will always be the case, but it could help the search significantly.

Resources such as the table of all 4-valent graphs up to 10 vertices as provided by Cantarella, Chapman and Mastin [155] would be a useful starting place. By working through every possible decoration of the graphs with over, under and virtual crossings, an exhaustive list could be made, and then analysed with the generalised Kauffman bracket. It would be difficult to guarantee that there are no pairs of distinct diagrams incorrectly recognised as equivalent, but it would at least be better than what is currently available.

Having this table would allow a more detailed analysis of the types of knots which appear in virtual closure. While most knots in proteins are simple, the random walks rapidly become unidentifiable. It would be interesting to look at how the probabilities of each knot type and their relative weights vary with degree of confinement as has been done in closed walks [72, 73, 156, 76].

This would also help understand the significance of knot globes, the maps of which closure directions result in which knot types. We understand some aspects of knot globes, such as how knot type changes across borders and how, unless the curve conformation is very particular, borders between classical and virtual knots are much more likely than between different classical knots under virtual closure. However, when it comes to analysing a curve we only look at the total areas of each knot type and neglect all of this additional spatial information. We have produced graphs showing how the knotted regions of the globe connect, such as in Fig 6.3, to try to capture some of this. Unfortunately, we don't really understand how best to use these graphs to understand the knotting of the curve. We know that the graphs of the virtual closure globe ought to be symmetric, owing to the mirror symmetry of the globe it is derived from, and all our previous comments about which knotted regions can border which apply to connected nodes here, but this is only a recasting of what we already know.

One step towards understanding the spatial properties of closure analyses, and indeed making the whole closure analysis more satisfying, would be to

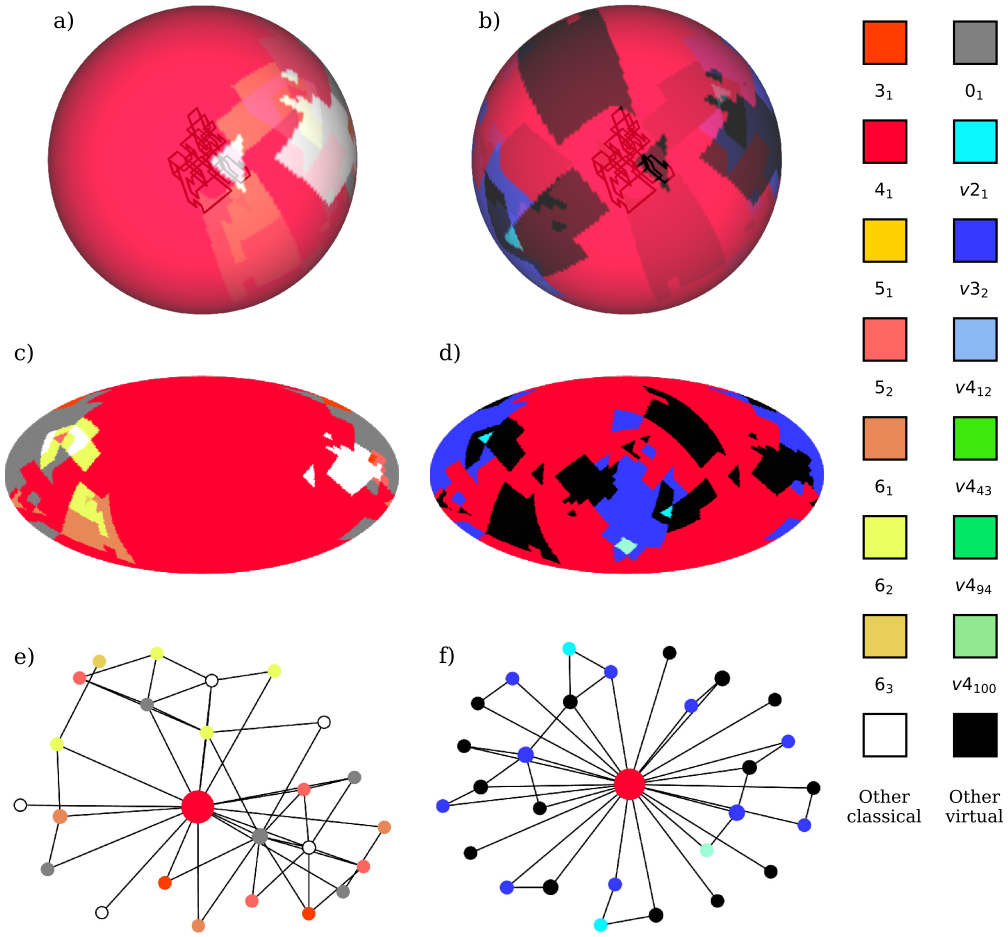


Figure 6.3: The knot globe, map and graph for a curve which is strongly figure-eight knotted under both sphere and virtual closure. a), c) and e) use sphere closure while b) d) and f) use virtual closure. There are some graphing errors in plots e) and f) due to the resolution available and how we determine the boundaries.

determine the complete, continuous knot globe, as opposed to the discrete one currently obtained from a finite number of sampled points. One could do this by projecting a line from an end-point to part of the rest of the curve, and extending this to the surface of the globe. By tracing along the length of the curve, all the possible borders due to the passing of this end-point across a strand in projection will be found. If the curve is piecewise linear, the arc between vertices of the curve will form sections of great circles on the knot globe. Performing the same procedure for the other end-point will give all the different borders between knot types, and many other borders over which the knot type does not change. One can then sample just a single closure within

each area marked by these borders and 'paint by numbers' to get the complete knot globe.

Something that has been assumed throughout has been the relative stability of strong and weak knots. Clearly, a curve with ends far extended from the knotted region will be fairly stable to small perturbations in the shape of the curve, provided crossing changes are prevented. If the ends are closer to the tangling, it is likely that small perturbations can change the knotting seen in virtual closure. The weaker the knotting, the more likely it is that a small change could considerably change the knotting. This could have physical impacts for dynamical systems, potentially allowing for easier unknotting, or knotting. A more quantitative understanding of this is lacking.

Finally, the extension to different sorts of objects such as links would be interesting. Virtual knots extend to virtual links just as classical knots extend to classical links and could be useful in analysing systems of multiple components. The question of how to close links in a sphere closure type approach is complicated as there are potential consequences for the linking depending on the closure points chosen. This can never be an issue with virtual closure as a single projection direction is all that can be chosen.

In proteins, we have neglected to include bonds other than peptide bonds between neighbouring amino acids, but bonds such as disulfide bridges between sequence distant amino acids are common. Treating these extra bonds as branch points, we can represent the protein backbone as an embedding of a more complex graph. There can be closed cycles in these graphs which can be understood using classical knot theory, or the whole structure can be understood as a theta curve. It is unclear how best to incorporate the open ends into such a structure. There are early attempts to use knotoids to address the problem with the introduction of bonded knotoids [125]. These require the transformation of a branch point into a link, understanding the final structure as a multi-component knotoid. Perhaps applying a virtual closure scheme to produce virtual theta curves could be a useful method to tackle this problem also.

We hope that the work presented here has convinced the reader that virtual knots offer a useful tool for the analysis of open curves. Additionally, the coverage of the most common knot and the distinction between strong and weak knotting we believe are important concepts when analysing open curve knotting, likely to be relevant in future studies. Hopefully the groundwork laid

here will give researchers the knowledge to look in-depth at the knotting of more geometrically complex and compact systems than before.

Appendices

Additional knot globes, maps and graphs

A

In this appendix we present a number of figures showing different representations of the knot globe for a variety of knotted curves. In each figure, a) will show the knot globe on sphere closure and b) on virtual closure. c) and d) are Mollweide projections of these respective globes. e) and f) will show graphs of the connected regions on the knot globe, with the size of each node giving a suggestion of the area of the region it represents. These are limited by the resolution of the closure analysis and how we determine the border position. In a perfect analysis, some separate nodes should be together, and some small nodes may be missed all together. It is possible to always represent these graphs in a planar fashion, although they are not always presented so here, the priority given to straight edges, minimal edge length and adequate node separation.

A.1 Lattice walks

All of these curves are walks on a $6 \times 6 \times 6$ cubic lattice, as in Chapter 5. Some of these curves were shown in Chapter 3, one for each of strong classical, strong virtual, weak classical, weak virtual and weak total knotting under virtual closure. We provide an additional example of each type of knotting here.

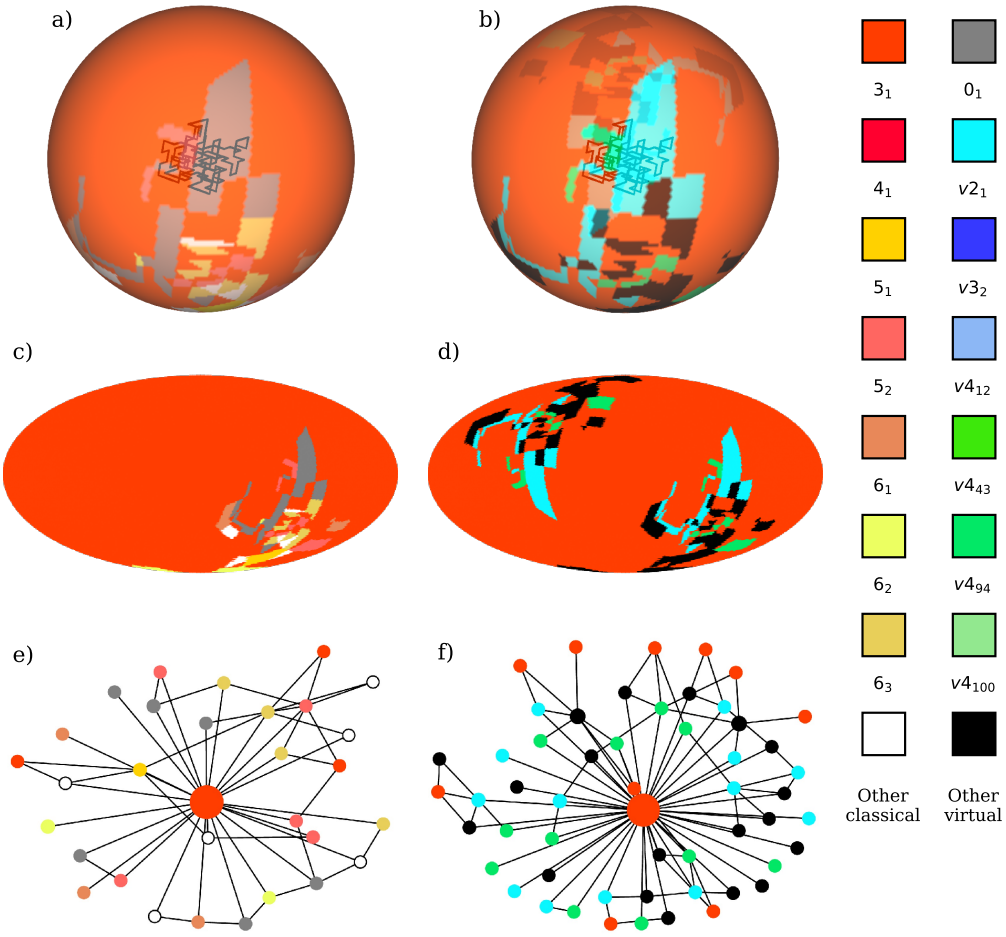


Figure A.1: This curve is strongly trefoil knotted under both sphere and virtual closure.

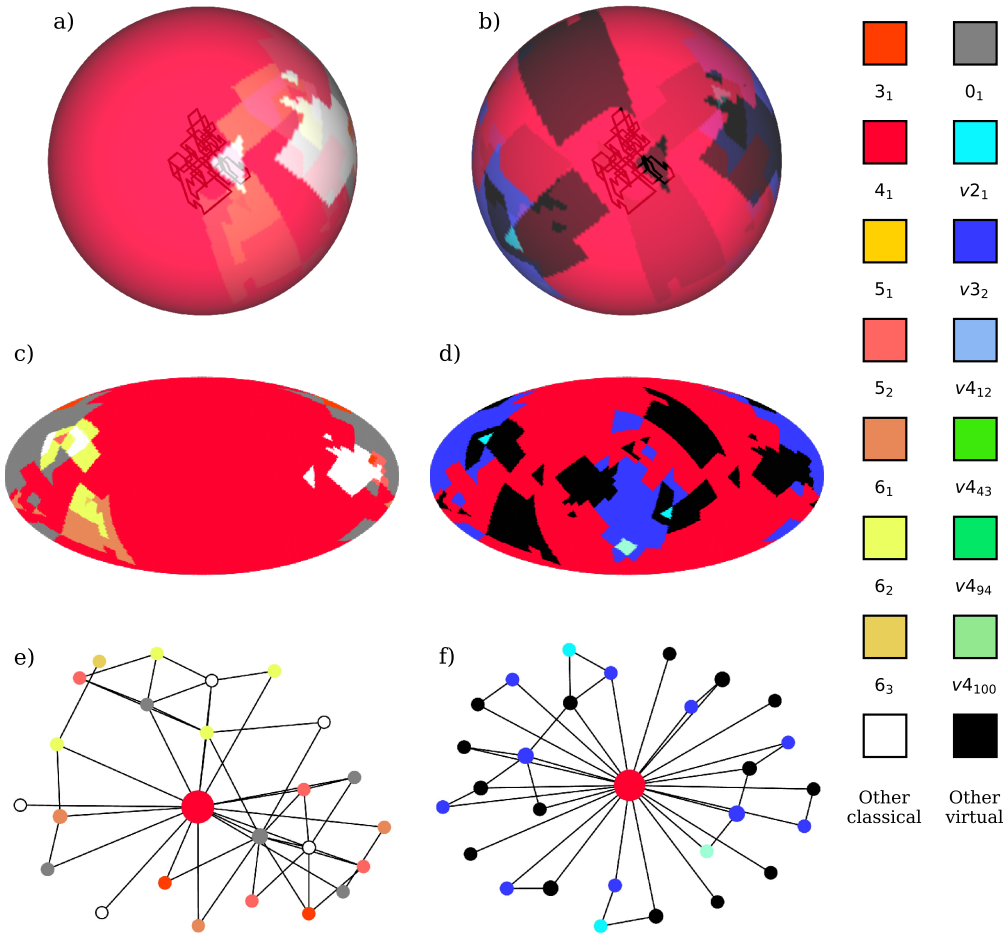


Figure A.2: This curve is strongly figure-eight knotted under both sphere and virtual closure. However, there is more competition between knot types here than in Fig A.1.

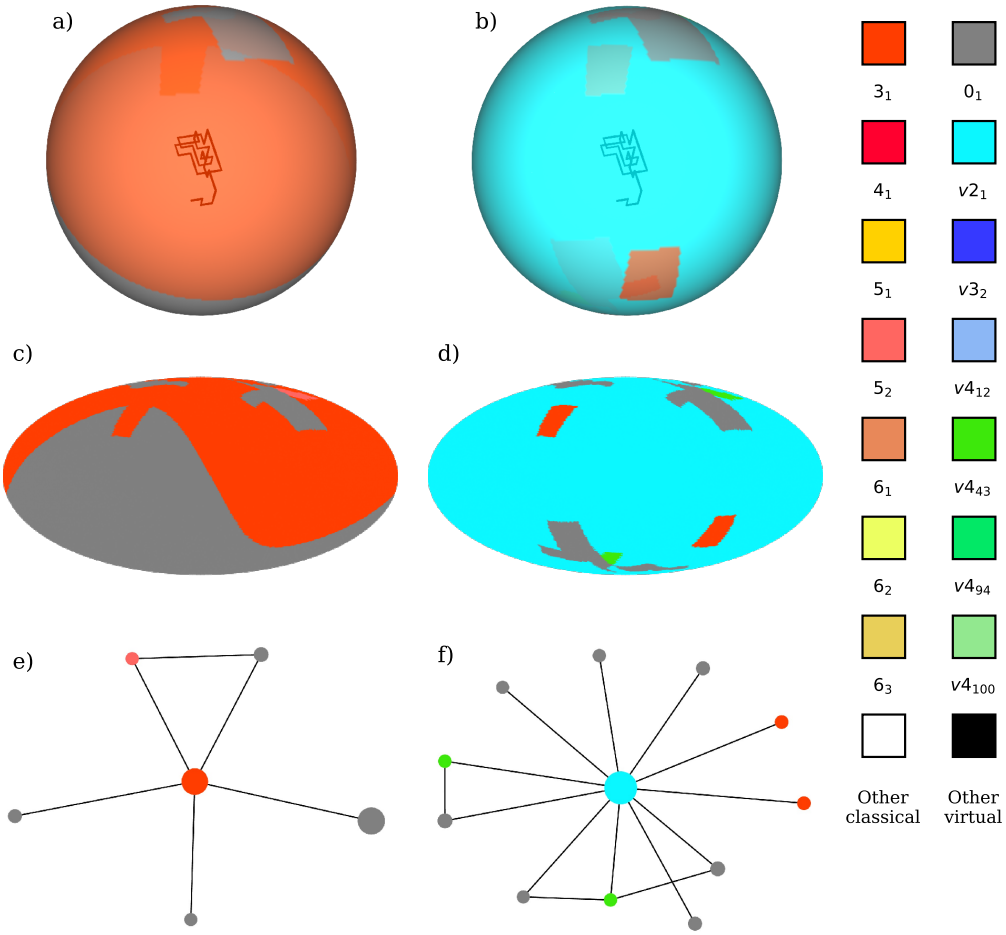


Figure A.3: This curve is unknotted under sphere closure, with a 52% coverage of the unknot, but strongly $v2_1$ knotted under virtual closure.

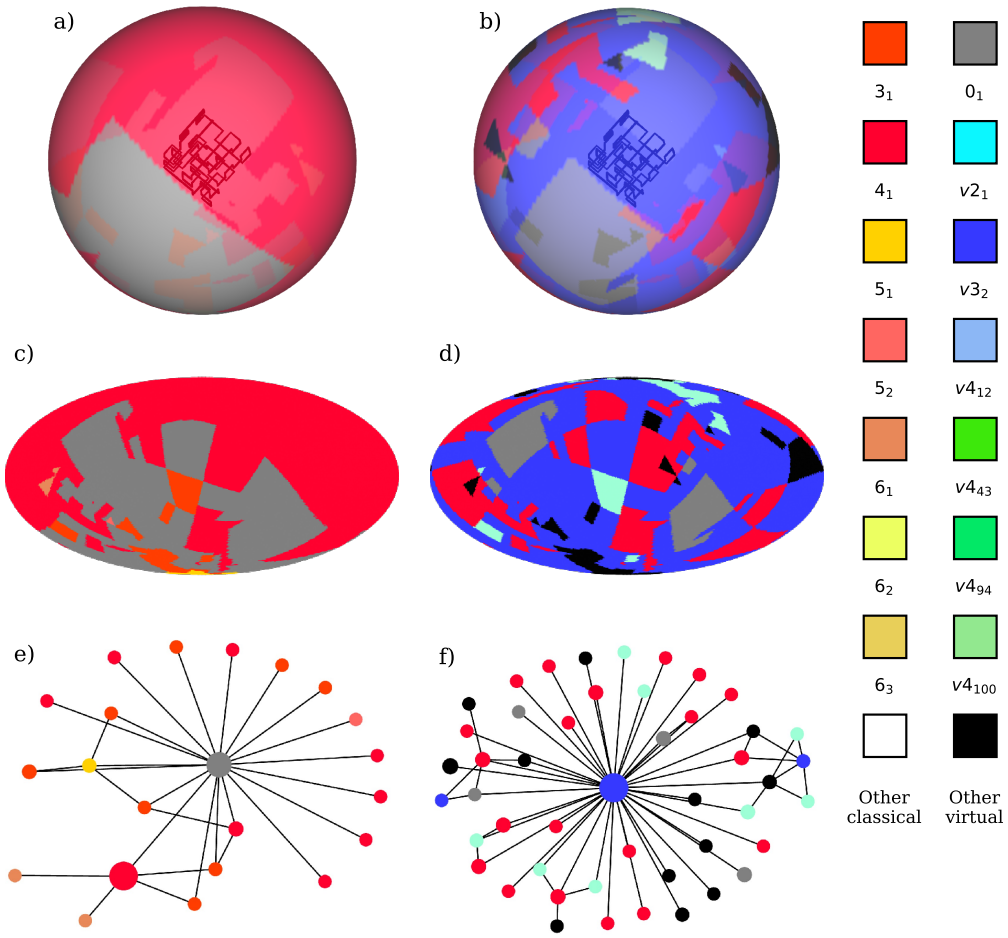


Figure A.4: This curve is strongly 4_1 knotted under sphere closure, and strongly $v3_2$ knotted under virtual closure. $v3_2$ does not dominate as much as $v2_1$ did in Fig A.3.

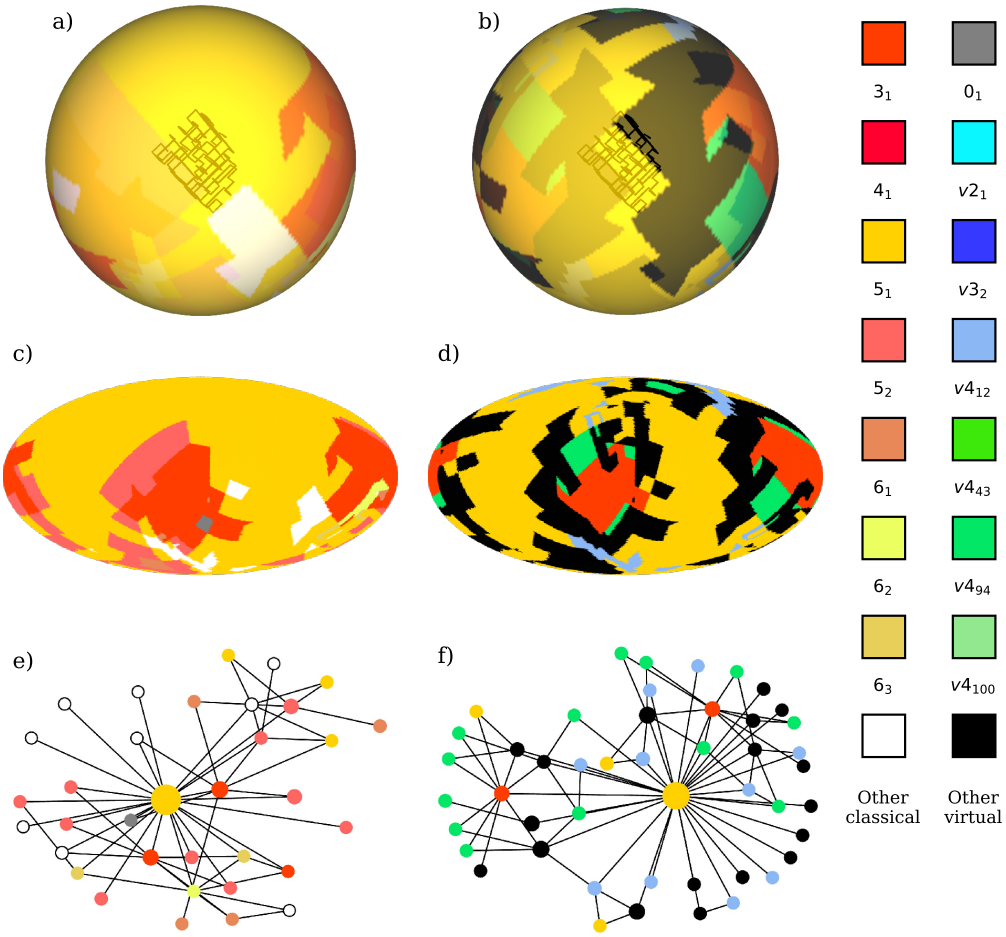


Figure A.5: This curve is strongly 5_1 knotted under sphere closure, but weakly classical knotted under virtual closure. The 5_1 areas under sphere closure have been eaten away by virtual knots under virtual closure, reducing its dominance to below 50%.

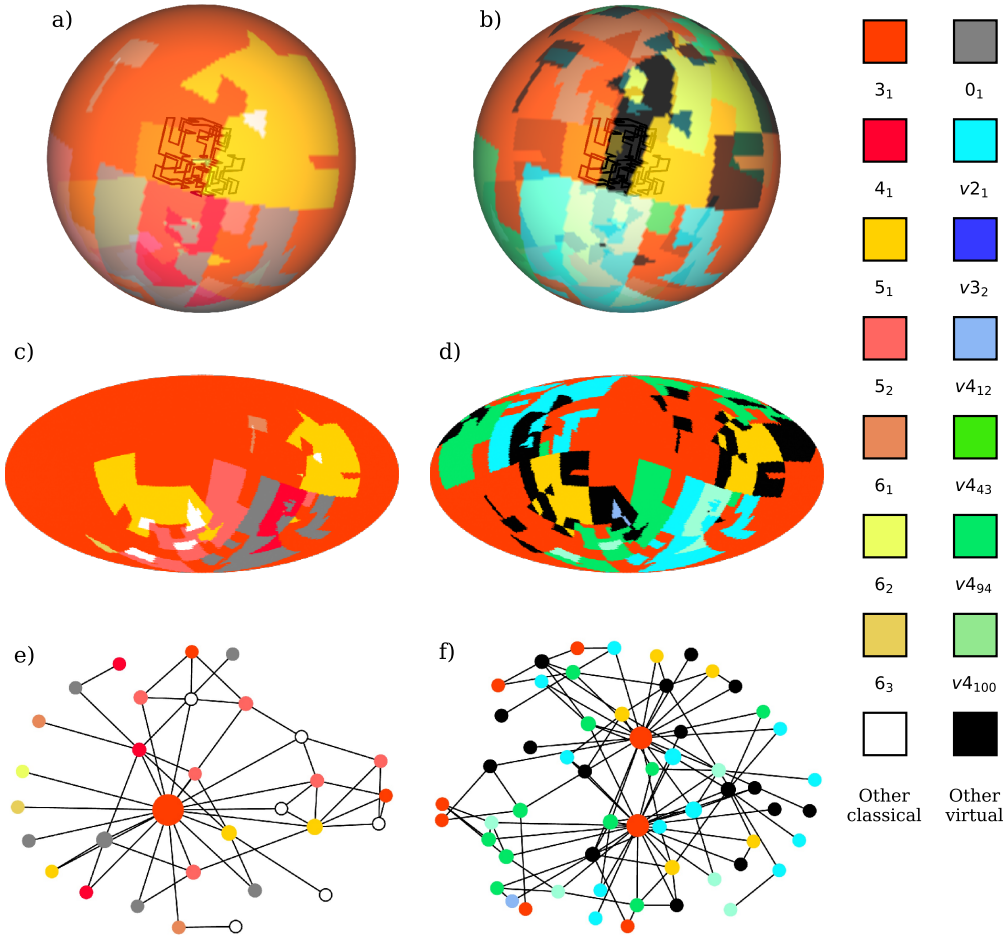


Figure A.6: Another curve strongly knotted under sphere closure and weakly classically knotted under virtual closure. Here the trefoil knot previously dominated.

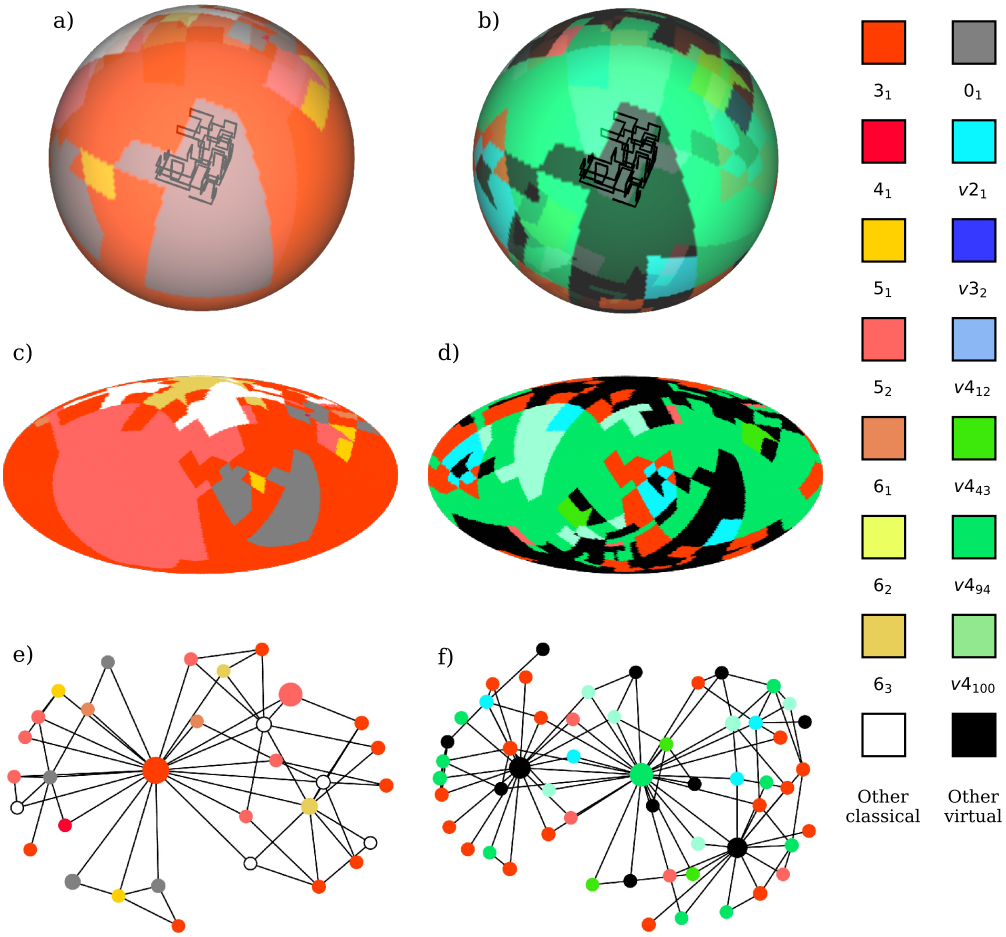


Figure A.7: This curve is weakly classically knotted under sphere closure, and weakly virtual knotted under virtual closure. No unknotted regions remain under virtual closure.

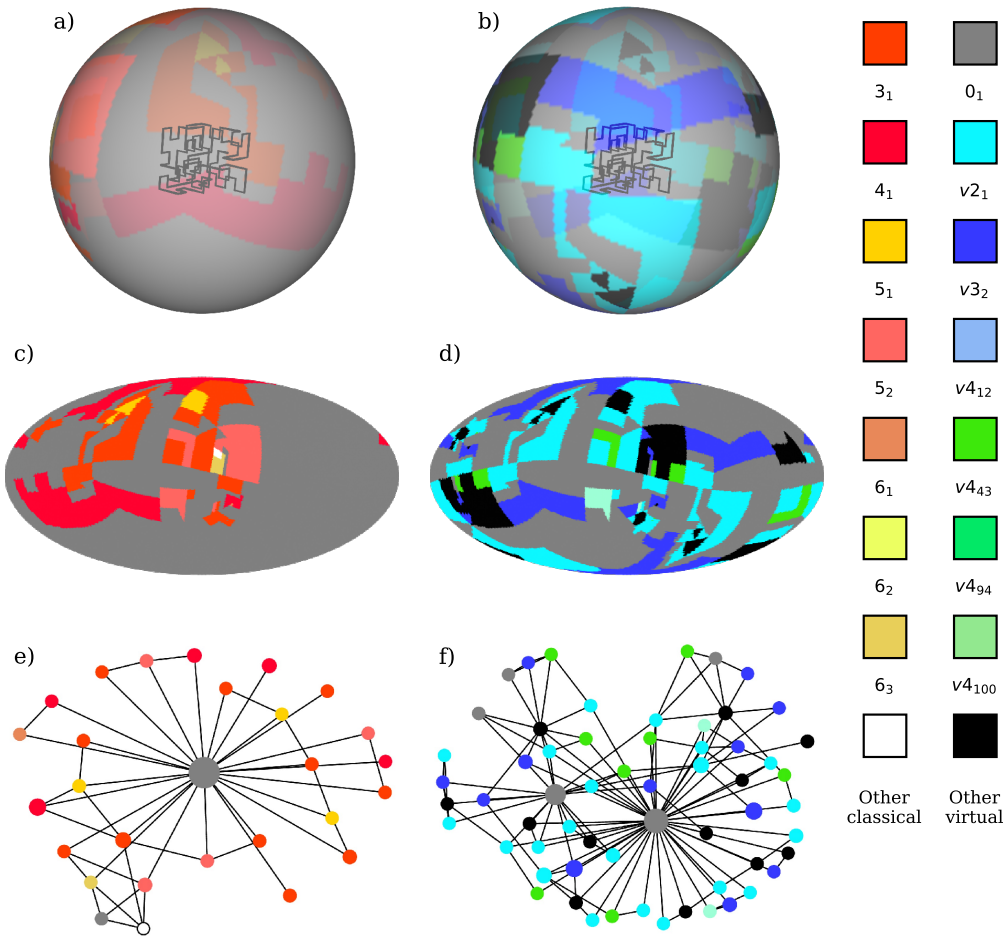


Figure A.8: A curve which is unknotted under sphere closure and weakly virtual knotted under virtual closure. Knots only cover 55% of closures in b) and d) so this curve is close to being unknotted.

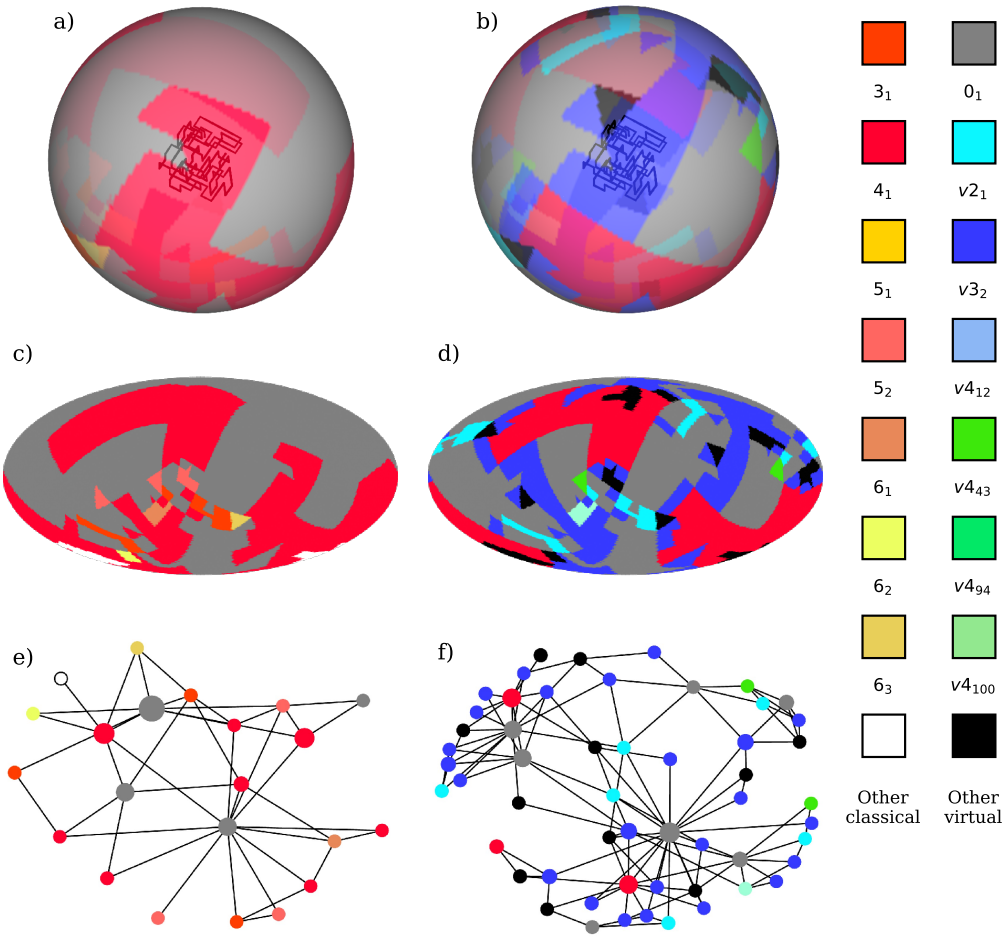


Figure A.9: An unknotted curve under sphere closure (52% unknot) which is weak total knotted under virtual closure. Unknots still make up 40% of virtual closures, with classicals comprising 25% and virtuals the remaining 35%.

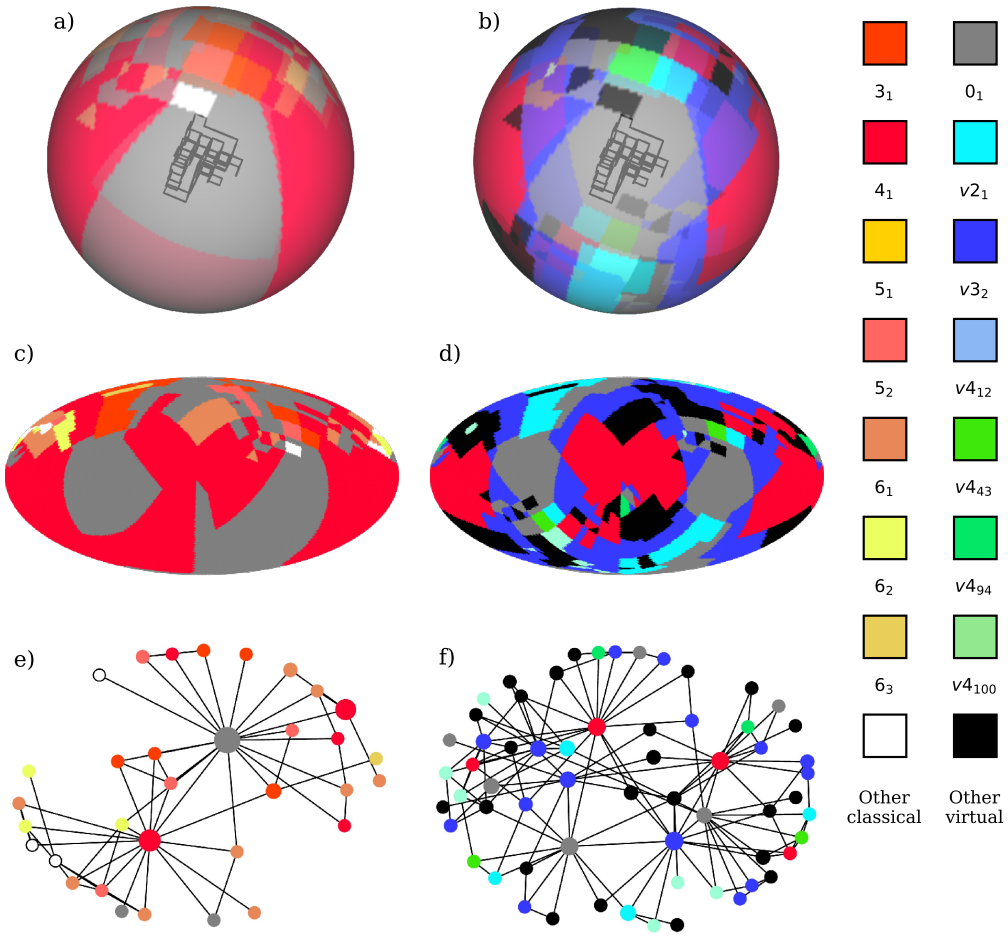


Figure A.10: A borderline weak knotted curve under sphere closure, with 49% 4_1, which is weakly total knotted under virtual closure. Here, unknots make up only 19% of virtual closures, with classicals covering 33% and virtuals 48%.

A.2 Proteins

The curves in this section are protein backbones. They are in order of appearance in the thesis, the first from Chapter 3 where both maps and globes were shown, and the others from Chapter 4 where only maps were shown.

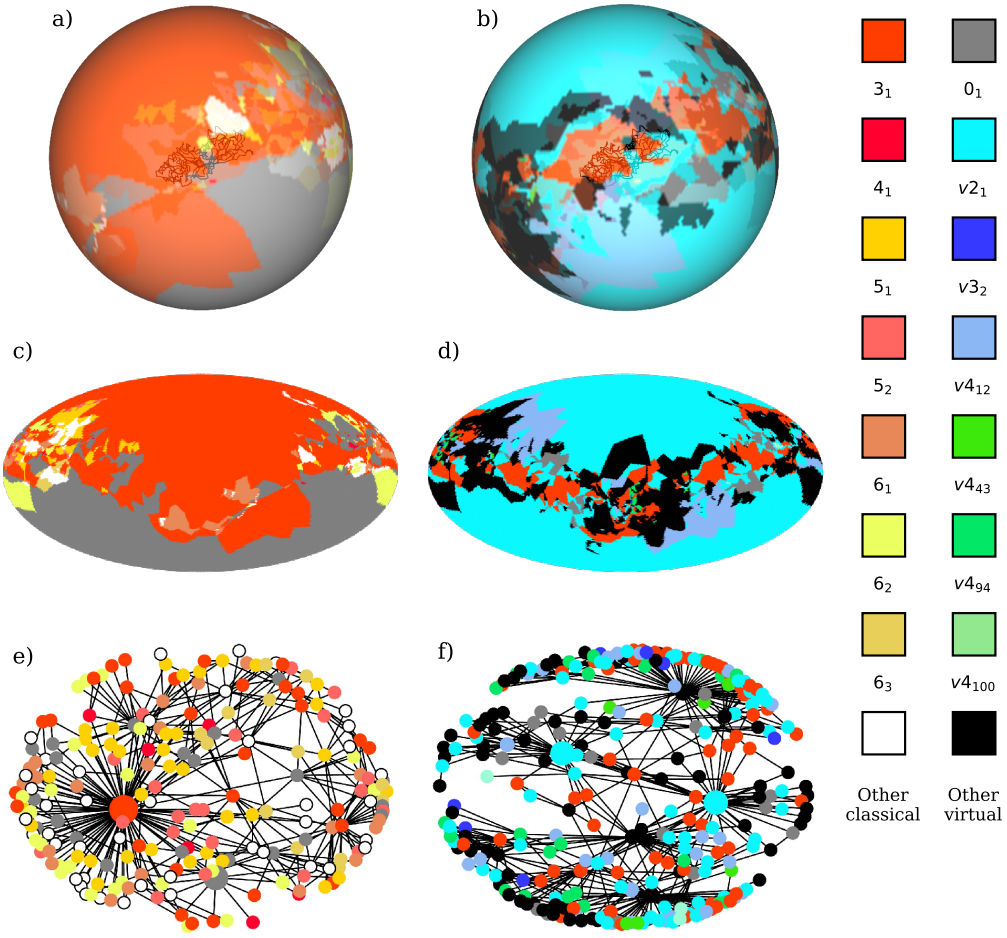


Figure A.11: This curve is strongly trefoil knotted under sphere closure and strongly $v2_1$ knotted under virtual closure.

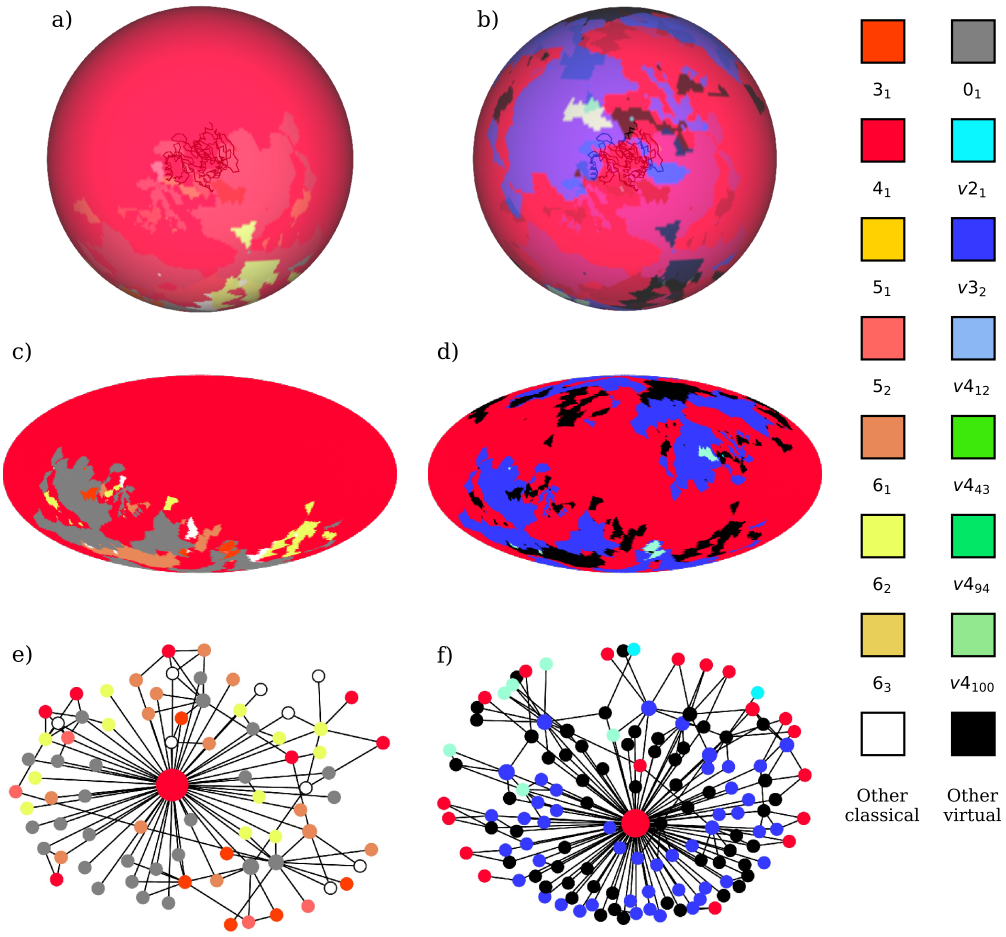


Figure A.12: PDB ID 4E04, chain A. Strongly 4_1 knotted under both sphere and virtual closure.

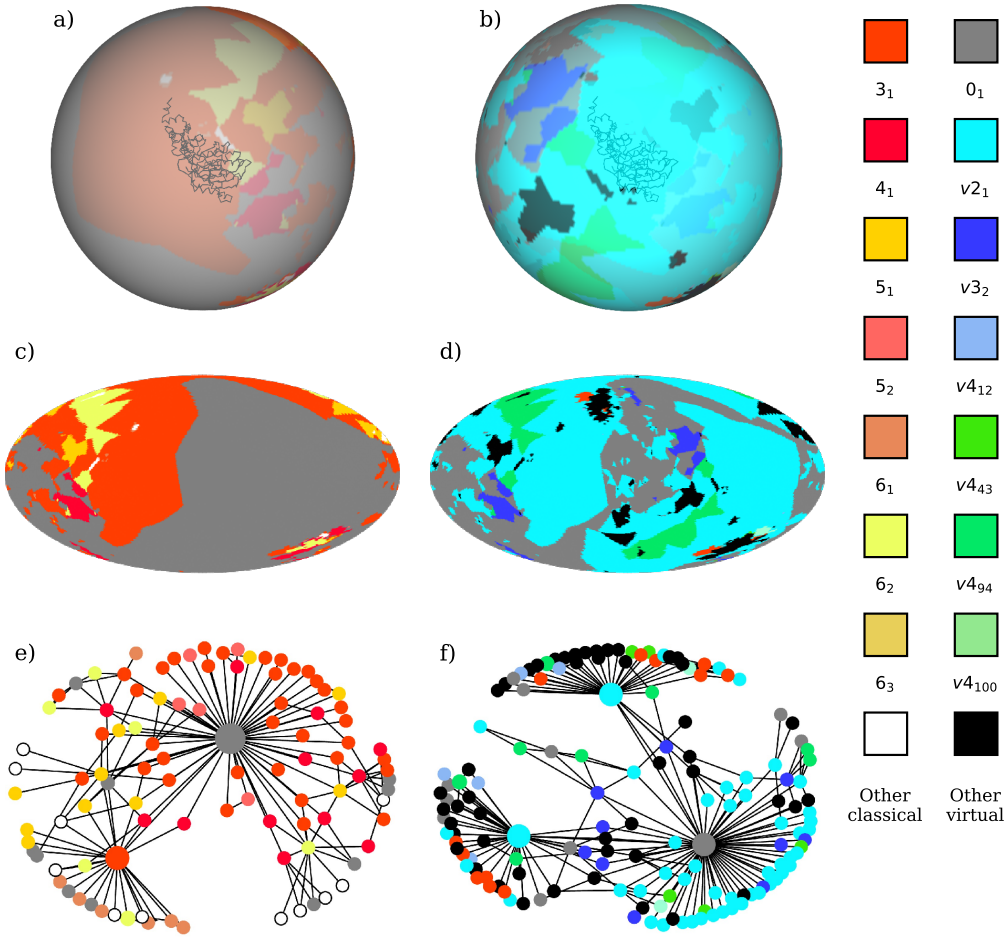


Figure A.13: PDB ID 3WKU, chain B. Unknotted under sphere closure, but strongly $v2_1$ knotted under virtual closure.

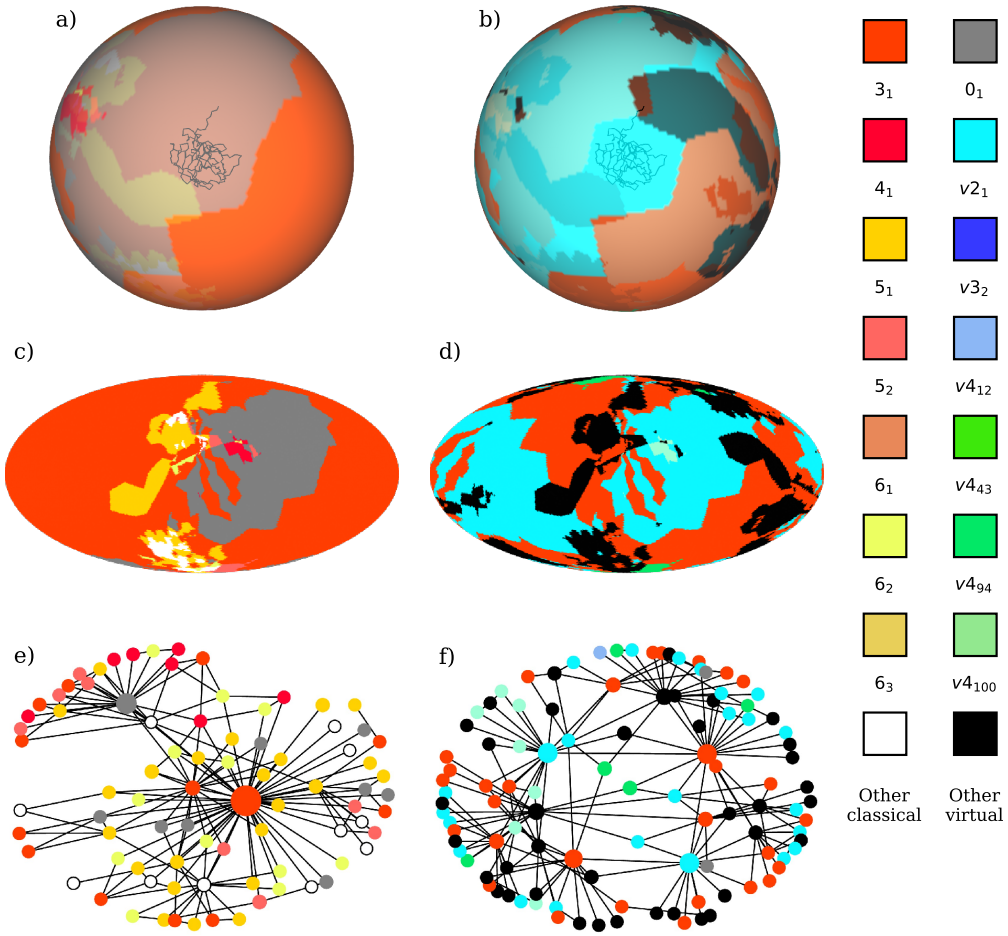


Figure A.14: PDB ID 4XIX, chain A. Strongly 3_1 knotted under sphere closure, weakly virtual knotted under virtual closure.

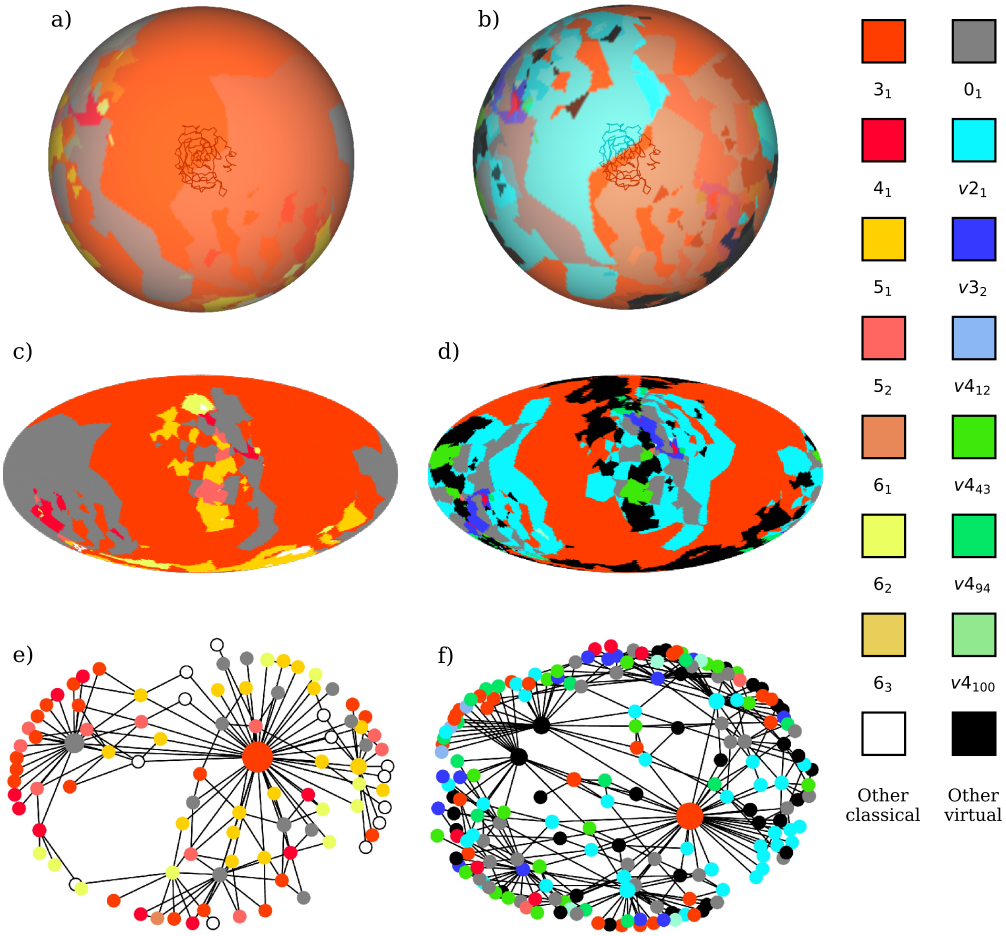


Figure A.15: PDB ID 3KIG, chain A. Strongly 3_1 knotted under sphere closure, weakly total knotted under virtual closure.

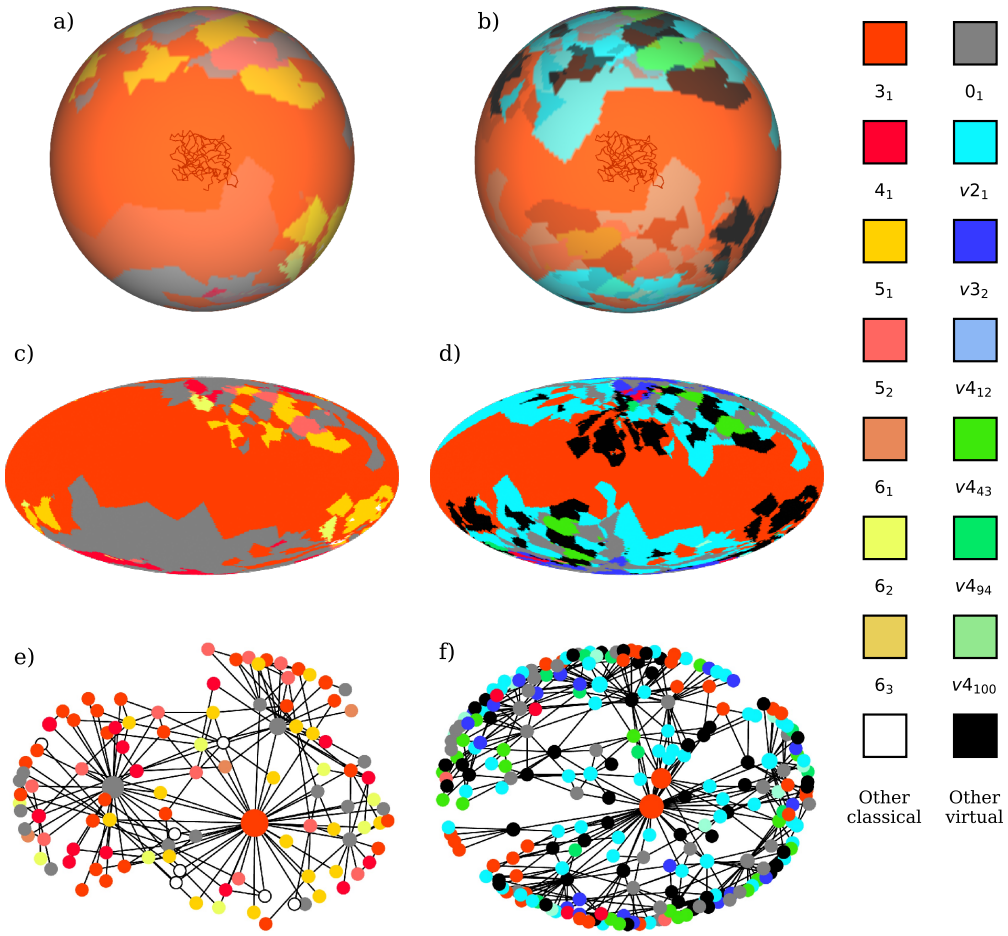


Figure A.16: PDB ID 1CZM, chain A. Strongly 3_1 knotted under sphere closure, weakly classical knotted under virtual closure.

Bibliography

[1] M Rubinstein and R H Colby. *Polymer physics*. Oxford University Press, (2003).

[2] Y Tezuka and H Oike. Topological polymer chemistry. *Progress in Polymer Science*, **27**:1069–1122, (2002).

[3] E Orlandini and S G Whittington. Statistical topology of closed curves: Some applications in polymer physics. *Reviews of Modern Physics*, **79**:611–642, (2007).

[4] M Kardar. The elusiveness of polymer knots. *The European Physical Journal B*, **64**:519–523, (2008).

[5] A Y Grosberg. A few notes about polymer knots. *Polymer Science Series A*, **51**:70–79, (2009).

[6] D J Mai and C M Schroeder. Single polymer dynamics of topologically complex DNA. *Current Opinion in Colloid & Interface Science*, **26**:28–40, (2016).

[7] E Orlandini. Statics and dynamics of DNA knotting. *Journal of Physics A: Mathematical and Theoretical*, **51**:053001, (2017).

[8] E Orlandini. Statistical topology and knotting of fluctuating filaments. *Physica A: Statistical Mechanics and its Applications*, **504**:155–175, (2017).

[9] H L Frisch and E Wasserman. Chemical topology. *Journal of the American Chemical Society*, **83**:3789–3795, (1961).

[10] M Delbrück. Knotting problems in biology. *Proceedings of Symposia in Applied Mathematics*, **14**:55–67, (1962).

[11] D W Sumners and S G Whittington. Knots in self-avoiding walks. *Journal of Physics A: Mathematical and General*, **21**:1689–1694, (1988).

[12] N Pippenger. Knots in random walks. *Discrete Applied Mathematics*, **25**:273–278, (1989).

[13] P J Flory. *Principles of Polymer Chemistry*. Cornell University Press, (1953).

[14] P Tompa. Intrinsically unstructured proteins. *Trends in Biochemical Sciences*, **27**:527–533, (2002).

[15] V N Uversky. Natively unfolded proteins: a point where biology waits for physics. *Protein Science*, **11**:739–756, (2002).

[16] H J Dyson and P E Wright. Intrinsically unstructured proteins and their functions. *Nature Reviews Molecular Cell Biology*, **6**:197–208, (2005).

[17] C I Branden and J Tooze. *Introduction to Protein Structure*, chapter 1. Garland Science, (1998).

[18] E Buxbaum. *Fundamentals of protein structure and function*, volume 31. Springer, (2007).

[19] S McNicholas, E Potterton, K S Wilson, and M E M Noble. Presenting your structures: the CCP4mg molecular-graphics software. *Acta Crystallographica Section D: Biological Crystallography*, **67**:386–394, (2011).

[20] M L Mansfield. Are there knots in proteins? *Nature Structural and Molecular Biology*, **1**:213–214, (1994).

[21] W R Taylor. A deeply knotted protein structure and how it might fold. *Nature*, **406**:916–919, (2000).

[22] R Lua, A L Borovinskiy, and A Y Grosberg. Fractal and statistical properties of large compact polymers: a computational study. *Polymer*, **45**:717–731, (2004).

[23] M Jamroz, W Niemyska, E J Rawdon, A Stasiak, K C Millett, P Sulkowski, and J L Sulkowska. Knotprot: a database of proteins with knots and slipknots. *Nucleic Acids Research*, **43**:D306–D314, (2014).

[24] C C Adams. *The Knot Book*. American Mathematical Society, (1994).

[25] B Marcone, E Orlandini, A L Stella, and F Zonta. What is the length of a knot in a polymer? *Journal of Physics A: Mathematical and General*, **38**:L15–L21, (2004).

[26] K Millett, A Dobay, and A Stasiak. Linear random knots and their scaling behavior. *Macromolecules*, **38**:601–606, (2005).

[27] K C Millett and B M Sheldon. Tying down open knots: a statistical method for identifying open knots with applications to proteins. In *Physical And Numerical Models In Knot Theory: Including Applications to the Life Sciences*, pages 203–217. World Scientific, (2005).

[28] L Tubiana, E Orlandini, and C Micheletti. Probing the entanglement and locating knots in ring polymers: a comparative study of different arc closure schemes. *Progress of Theoretical Physics Supplement*, **191**:192–204, (2011).

[29] L H Kauffman. Virtual knot theory. *European Journal of Combinatorics*, **20**:663–690, (1999).

[30] S Labeit, D Labeit, and H Granzier. Titin gene (TTN): description of the

gene coding for titin, a giant protein of critical importance for myofibrillar integrity and elasticity in vertebrate striated muscle. *eLS*, pages 1–6, (2018).

[31] J W Neidigh, R M Fesinmeyer, and N H Andersen. Designing a 20-residue protein. *Nature Structural and Molecular Biology*, **9**:425–430, (2002).

[32] OpenStax College. Biology. http://cnx.org/content/m44402/latest/?collection=col11448/latest, (2018).

[33] M Goedert. Alzheimer's and Parkinson's diseases: the prion concept in relation to assembled Aβ, tau, and α-synuclein. *Science*, **349**:1255555, (2015).

[34] J S Richardson. β-sheet topology and the relatedness of proteins. *Nature*, **268**:495–500, (1977).

[35] T N Bryant, H C Watson, and P L Wendell. Structure of yeast phosphoglycerate kinase. *Nature*, **247**:14–17, (1974).

[36] M L Mansfield. Fit to be tied. *Nature Structural and Molecular Biology*, **4**:166–167, (1997).

[37] G Kolesov, P Virnau, M Kardar, and L A Mirny. Protein knot server: detection of knots in protein structures. *Nucleic Acids Research*, **35**:W425–W428, (2007).

[38] Y L Lai, C C Chen, and J K Hwang. pKNOT: the protein KNOT web server. *Nucleic Acids Research*, **35**:W420–W424, (2007).

[39] J L Sulkowska, E J Rawdon, K C Millett, J N Onuchic, and A Stasiak. Conservation of complex knotting and slipknotting patterns in proteins. *Proceedings of the National Academy of Sciences*, **109**:E1715–E1723, (2012).

[40] W R Taylor. Protein knots and fold complexity: some new twists. *Computational Biology and Chemistry*, **31**:151–162, (2007).

[41] A L Mallam and S E Jackson. Knot formation in newly translated proteins is spontaneous and accelerated by chaperonins. *Nature Chemical Biology*, **8**:147–53, (2012).

[42] N C H Lim and S E Jackson. Mechanistic insights into the folding of knotted proteins in vitro and in vivo. *Journal of Molecular Biology*, **427**:248–258, (2015).

[43] A L Mallam, J M Rogers, and S E Jackson. Experimental detection of knotted conformations in denatured proteins. *Proceedings of the National Academy of Sciences*, **107**:8189–8194, (2010).

[44] T O Yeates, T S Norcross, and N P King. Knotted and topologically complex

proteins as models for studying folding and stability. *Current Opinion in Chemical Biology*, **11**:595–603, (2007).

[45] N P King, E O Yeates, and T O Yeates. Identification of rare slipknots in proteins and their implications for stability and folding. *Journal of Molecular Biology*, **373**:153–166, (2007).

[46] N P King, A W Jacobitz, M R Sawaya, L Goldschmidt, and T O Yeates. Structure and folding of a designed knotted protein. *Proceedings of the National Academy of Sciences*, **107**:20732–20737, (2010).

[47] T C Sayre, T M Lee, N P King, and T O Yeates. Protein stabilization in a highly knotted protein polymer. *Protein Engineering, Design & Selection*, **24**:627–630, (2011).

[48] P Dabrowski-Tumanski and J I Sulkowska. To tie or not to tie? that is the question. *Polymers*, **9**:454, (2017).

[49] P Dabrowski-Tumanski, A Stasiak, and J I Sulkowska. In search of functional advantages of knots in proteins. *PLoS ONE*, **11**:e0165986, (2016).

[50] E Flapan and G Heller. Topological complexity in protein structures. *Molecular Based Mathematical Biology*, **3**:23–42, (2015).

[51] P Dabrowski-Tumanski, A I Jarmolinska, W Niemyska, E J Rawdon, K C Millett, and J I Sulkowska. Linkprot: A database collecting information about biological links. *Nucleic Acids Research*, **45**:gkw976, (2016).

[52] P Dabrowski-Tumanski and J I Sulkowska. Topological knots and links in proteins. *Proceedings of the National Academy of Sciences*, **114**:3415–3420, (2017).

[53] E Haglund, J I Sulkowska, J K Noel, H Lammert, J N Onuchic, and P A Jennings. Pierced lasso bundles are a new class of knot-like motifs. *PLoS Computational Biology*, **10**:e1003613, (2014).

[54] W Niemyska, P Dabrowski-Tumanski, M Kadlof, E Haglund, P Sułkowski, and J I Sulkowska. Complex lasso: new entangled motifs in proteins. *Scientific Reports*, **6**:36895, (2016).

[55] P Dabrowski-Tumansk, W Niemyska, P Pasznik, and J I Sulkowska. Lassoprot: server to analyze biopolymers with lassos. *Nucleic Acids Research*, **44**:W383–W389, (2016).

[56] W R Wikoff, L Liljas, R L Duda, H Tsuruta, R W Hendrix, and J E Johnson. Topologically linked protein rings in the bacteriophage hk97 capsid. *Science*, **289**:2129–2133, (2000).

[57] V V Rybenkov, C Ullsperger, A V Vologodskii, and N R Cozzarelli. Sim-

plification of DNA topology below equilibrium values by type II topoisomerases. *Science*, **277**:690–693, (1997).

[58] G R Buck and E L Zechiedrich. DNA disentangling by type-2 topoisomerases. *Journal of Molecular Biology*, **340**:933–939, (2004).

[59] M K Shimamura and T Deguchi. Finite-size and asymptotic behaviors of the gyration radius of knotted cylindrical self-avoiding polygons. *Physical Review E*, **65**:051802, (2002).

[60] S R Quake. Topological effects of knots in polymers. *Physical Review Letters*, **73**:3317–3320, (1994).

[61] Y Zhao and F Ferrari. Topological effects on the mechanical properties of polymer knots. *Physica A: Statistical Mechanics and its Applications*, **486**:44–64, (2017).

[62] B W Soh, V Narsimhan, A R Klotz, and P S Doyle. Knots modify the coil–stretch transition in linear DNA polymers. *Soft Matter*, **14**:1689–1698, (2018).

[63] R Matthews, A A Louis, and J M Yeomans. Knot-controlled ejection of a polymer from a virus capsid. *Physical Review Letters*, **102**:088101, (2009).

[64] P Grassberger. Opacity and entanglement of polymer chains. *Journal of Physics A: Mathematical and General*, **34**:9959, (2001).

[65] A V Vologodskii, A V Lukashin, M D Frank-Kamenetskii, and V V Anshelevich. The knot problem in statistical mechanics of polymer chains. *Soviet Journal of Experimental and Theoretical Physics*, **39**:1059–1063, (1974).

[66] W S Kendall. The knotting of brownian motion in 3-space. *Journal of the London Mathematical Society*, **2**:378–384, (1979).

[67] K Koniaris and M Muthukumar. Knottedness in ring polymers. *Physical Review Letters*, **66**:2211–2214, (1991).

[68] K Koniaris and M Muthukumar. Self-entanglement in ring polymers. *The Journal of Chemical Physics*, **95**:2873–2881, (1991).

[69] N T Moore and A Y Grosberg. The abundance of unknots in various models of polymer loops. *Journal of Physics A: Mathematical and General*, **39**:9081–9092, (2006).

[70] Y Diao, N Pippenger, and D W Sumners. On random knots. In *Random knotting and linking*, pages 187–197. World Scientific, (1994).

[71] Y Diao. The knotting of equilateral polygons in R^3. *Journal of Knot Theory and Its Ramifications*, **4**:189–196, (1995).

[72] T Deguchi and K Tsurusaki. Topology of closed random polygons. *Journal*

of the Physical Society of Japan, **62**:1411–1414, (1993).

[73] T Deguchi and K Tsurusaki. A statistical study of random knotting using the vassiliev invariants. *Journal of Knot Theory and Its Ramifications*, **3**:321–353, (1994).

[74] T Deguchi and K Tsurusaki. Universality of random knotting. *Physical Review E*, **55**:6245–6248, (1997).

[75] M Baiesi, E Orlandini, and A L Stella. Ranking knots of random, globular polymer rings. *Physical Review Letters*, **99**:058301, (2007).

[76] E Uehara and T Deguchi. Knotting probability of self-avoiding polygons under a topological constraint. *The Journal of Chemical Physics*, **147**:094901, (2017).

[77] J des Cloizeaux. Ring polymers in solution: topological effects. *Journal de Physique Lettres*, **42**:433–436, (1981).

[78] A Dobay, J Dubochet, K Millett, P E Sottas, and A Stasiak. Scaling behavior of random knots. *Proceedings of the National Academy of Sciences*, **100**:5611–5615, (2003).

[79] N T Moore, R C Lua, and A Y Grosberg. Topologically driven swelling of a polymer loop. *Proceedings of the National Academy of Sciences of the United States of America*, **101**:13431–13435, (2004).

[80] J Suzuki, A Takano, and Y Matsushita. Topological constraint in ring polymers under theta conditions studied by monte carlo simulation. *The Journal of Chemical Physics*, **138**:024902, (2013).

[81] E Uehara and T Deguchi. Scaling behavior of knotted random polygons and self-avoiding polygons: Topological swelling with enhanced exponent. *The Journal of Chemical Physics*, **147**:214901, (2017).

[82] A Y Grosberg, A Feigel, and Y Rabin. Flory-type theory of a knotted ring polymer. *Physical Review E*, **54**:6618–6622, (1996).

[83] X R Bao, H J Lee, and S R Quake. Behavior of complex knots in single DNA molecules. *Physical Review Letters*, **91**:265506, (2003).

[84] E Orlandini, A L Stella, and C. Vanderzande. Loose, flat knots in collapsed polymers. *Journal of Statistical Physics*, **115**:681–700, (2004).

[85] P Virnau, Y Kantor, and M Kardar. Knots in globule and coil phases of a model polyethylene. *Journal of the American Chemical Society*, **127**:15102–15106, (2005).

[86] B Marcone, E Orlandini, A L Stella, and F Zonta. Size of knots in ring polymers. *Physical Review E*, **75**:041105, (2007).

[87] L Tubiana, E Orlandini, and C Micheletti. Multiscale entanglement in ring polymers under spherical confinement. *Physical Review Letters*, **107**:188302, (2011).

[88] A Y Grosberg. Do knots self-tighten for entropic reasons? *Polymer Science Series A*, **58**:864–872, (2016).

[89] G Gregoriadis. Engineering liposomes for drug delivery: progress and problems. *Trends in Biotechnology*, **13**:527–537, (1995).

[90] J Arsuaga, M Vázquez, S Trigueros, D W Sumners, and J Roca. Knotting probability of DNA molecules confined in restricted volumes: DNA knotting in phage capsids. *Proceedings of the National Academy of Sciences*, **99**:5373–5377, (2002).

[91] J Arsuaga, M Vazquez, P McGuirk, S Trigueros, D W Sumners, and J Roca. DNA knots reveal a chiral organization of DNA in phage capsids. *Proceedings of the National Academy of Sciences of the United States of America*, **102**:9165–9169, (2005).

[92] C Micheletti, D Marenduzzo, E Orlandini, and D W Sumners. Simulations of knotting in confined circular DNA. *Biophysical Journal*, **95**:3591–3599, (2008).

[93] D Marenduzzo, E Orlandini, A Stasiak, D W Sumners, L Tubiana, and C Micheletti. DNA-DNA interactions in bacteriophage capsids are responsible for the observed DNA knotting. *Proceedings of the National Academy of Sciences*, **106**:22269–22274, (2009).

[94] D Marenduzzo, C Micheletti, E Orlandini, and D W Sumners. Topological friction strongly affects viral DNA ejection. *Proceedings of the National Academy of Sciences*, **110**:20081–20086, (2013).

[95] P Poier, C N Likos, and R Matthews. Influence of rigidity and knot complexity on the knotting of confined polymers. *Macromolecules*, **47**:3394–3400, (2014).

[96] J P J Michels and F W Wiegel. Probability of knots in a polymer ring. *Physics Letters A*, **90**:381–384, (1982).

[97] E J J Van Rensburg and S G Whittington. The knot probability in lattice polygons. *Journal of Physics A: Mathematical and General*, **23**:3573–3590, (1990).

[98] C Micheletti, D Marenduzzo, E Orlandini, and D W Summers. Knotting of random ring polymers in confined spaces. *The Journal of Chemical Physics*, **124**:124–133, (2006).

[99] M Lautout-Magat. Contribution to the study of self-avoiding random walks (sarw) confined to strips and capillaries. *Journal of Polymer Science Part A: Polymer Chemistry*, **20**:2705–2713, (1982).

[100] C Micheletti and E Orlandini. Knotting and metric scaling properties of DNA confined in nano-channels: a monte carlo study. *Soft Matter*, **8**:10959–10968, (2012).

[101] E Orlandini and C Micheletti. Knotting of linear DNA in nano-slits and nano-channels: a numerical study. *Journal of Biological Physics*, **39**:267–275, (2013).

[102] M C Tesi, E J J van Rensburgs, E Orlandini, and S G Whittington. Knot probability for lattice polygons in confined geometries. *Journal of Physics A: Mathematical and General*, **27**:347–360, (1994).

[103] H P Hsu and P Grassberger. Polymers confined between two parallel plane walls. *The Journal of Chemical Physics*, **120**:2034–2041, (2004).

[104] L Dai, J R C van der Maarel, and P S Doyle. Effect of nanoslit confinement on the knotting probability of circular DNA. *ACS Macro Letters*, **1**:732–736, (2012).

[105] C Micheletti and E Orlandini. Numerical study of linear and circular model DNA chains confined in a slit: metric and topological properties. *Macromolecules*, **45**:2113–2121, (2012).

[106] J Cantarella, K Chapman, P Reiter, and C Shonkwiler. Open and closed random walks with fixed edgelengths in R^d. *arXiv preprint arXiv:1806.00079*, (2018).

[107] J Hoste, M Thistlethwaite, and J Weeks. The first $1,701,936$ knots. *The Mathematical Intelligencer*, **20**:33–48, (1998).

[108] J W Alexander. Topological invariants of knots and links. *Transactions of the American Mathematical Society*, **30**:275–306, (1928).

[109] V F R Jones. A polynomial invariant for knots and links via Von Neumann algebras. *Bulletin of the American Mathematical Society*, **12**:103–111, (1985).

[110] P Freyd, D Yetter, J Hoste, W B R Lickorish, K Millett, and A Ocneanu. A new polynomial invariant of knots and links. *Bulletin of the American Mathematical Society*, **12**:239–246, (1985).

[111] J H Przytycki and P Traczyk. Invariants of links of Conway type. *Kobe Journal of Mathematics*, **4**:115–139, (1987).

[112] J Green and D Bar-Natan. A table of virtual knots.

https://www.math.toronto.edu/drorbn/Students/GreenJ/, (2018). last updated Aug 04.

[113] K Alexander, A J Taylor, and M R Dennis. Proteins analysed as virtual knots. *Scientific Reports*, **7**:42300, (2017).

[114] A A Andreevna and S V Matveev. Classification of genus 1 virtual knots having at most five classical crossings. *Journal of Knot Theory and Its Ramifications*, **23**:1450031, (2014).

[115] L H Kauffman. A self-linking invariant of virtual knots. *arXiv preprint math/0405049*, (2004).

[116] L H Kauffman and D E Radford. Bioriented quantum algebras and a generalized Alexander polynomial for virtual links. In *Diagrammatic Morphisms and Applications*, volume 318 of *Contemporary Mathematics*, pages 113–140. American Mathematical Society, (2003).

[117] J Sawollek. On Alexander-Conway polynomials for virtual knots and links. arXiv:math/9912173v2, (1999).

[118] K C Millett, E J Rawdon, A Stasiak, and J L Sulkowska. Identifying knots in proteins. *Biochemical Society Transactions*, **41**:533–537, (2013).

[119] S F Edwards. The theory of rubber elasticity. *Polymer International*, **9**:140–143, (1977).

[120] E. A. Rakhmanov, E B Saff, and Y M Zhou. Minimal discrete energy on the sphere. *Mathematical Research Letters*, **1**:647–662, (1994).

[121] R Benlloch, D Shevela, T Hainzl, C Grundström, T Shutova, J Messinger, G Samuelsson, and A E Sauer-Eriksson. Crystal structure and functional characterization of photosystem ii-associated carbonic anhydrase cah3 in chlamydomonas reinhardtii. *Plant Physiology*, **167**:950–962, (2015).

[122] Wikipedia contributors. Mollweide projection. https://en.wikipedia.org/w/index.php?title=Mollweide_projection&oldid=826933376, (2018).

[123] F Wang, S Singh, J Zhang, T D Huber, K E Helmich, M Sunkara, K A Hurley, R D Goff, C A Bingman, A J Morris, J S Thorson, and G N Jr. Phillips. Understanding molecular recognition of promiscuity of thermophilic methionine adenosyltransferase sMAT from Sulfolobus solfataricus. *FEBS Journal*, **281**:4224–4239, (2014).

[124] D Goundaroulis, J Dorier, F Benedetti, and A Stasiak. Studies of global and local entanglements of individual protein chains using the concept of knotoids. *Scientific Reports*, **7**:6309, (2017).

[125] D Goundaroulis, N Gügümcü, S Lambropoulou, J Dorier, A Stasiak, and L Kauffman. Topological models for open-knotted protein chains using the concepts of knotoids and bonded knotoids. *Polymers*, **9**:444, (2017).

[126] H M Berman, J Westbrook, Z Feng, G Gilliland, T N Bhat, H Weissig, I N Shindyalov, and P E Bourne. The Protein Data Bank. *Nucleic Acids Research*, **28**:235–242, (2000).

[127] H Steen and M Mann. The abc's (and xyz's) of peptide sequencing. *Nature Reviews Molecular Cell Biology*, **5**:699–711, (2004).

[128] J C Kendrew, G Bodo, H M Dintzis, R G Parrish, and H Wyckoff. A three-dimensional model of the myoglobin molecule obtained by X-ray analysis. *Nature*, **181**:662–666, (1958).

[129] Y Shi. A glimpse of structural biology through X-ray crystallography. *Cell*, **159**:995–1014, (2014).

[130] A M Seddon, P Curnow, and P J Booth. Membrane proteins, lipids and detergents: not just a soap opera. *Biochimica et Biophysica Acta (BBA)–Biomembranes*, **1666**:105–117, (2004).

[131] E P Carpenter, K Beis, A D Cameron, and S Iwata. Overcoming the challenges of membrane protein crystallography. *Current Opinion in Structural Biology*, **18**:581–586, (2008).

[132] Z Huang and K-J Kim. Review of X-ray free-electron laser theory. *Physical Review Special Topics – Accelerators and Beams*, **10**:034801, (2007).

[133] R Neutze, G Brändén, and G F X Schertler. Membrane protein structural biology using X-ray free electron lasers. *Current Opinion in Structural Biology*, **33**:115–125, (2015).

[134] J L Smith, R F Fischetti, and M Yamamoto. Micro-crystallography comes of age. *Current Opinion in Structural Biology*, **22**:602–612, (2012).

[135] J Tenboer, S Basu, N Zatsepin, K Pande, D Milathianaki, M Frank, M Hunter, S Boutet, G J Williams, J E Koglin, D Oberthuer, M Heymann, C Kupitz, C Conrad, J Coe, S Roy-Chowdhury, U Weierstall, D James, D Wang, T Grant, A Barty, O Yefanov, J Scales, C Gati, C Seuring, V Srajer, R Henning, P Schwander, R Fromme, A Ourmazd, K Moffat, J J Van Thor, J C H Spence, P Fromme, H N Chapman, and M Schmidt. Time-resolved serial crystallography captures high-resolution intermediates of photoactive yellow protein. *Science*, **346**:1242–1246, (2014).

[136] J Cavanagh, W J Fairbrother, A G Palmer III, and N J Skelton. *Protein NMR spectroscopy: principles and practice*. Elsevier, (1995).

[137] J M Berg, J L Tymoczko, and L Stryer. *Biochemistry*. W.H.Freeman & Co Ltd, (2002).

[138] D Marion, P C Driscoll, L E Kay, P T Wingfield, A Bax, A M Gronenborn, and G M Clore. Overcoming the overlap problem in the assignment of proton NMR spectra of larger proteins by use of three-dimensional heteronuclear proton-nitrogen-15 Hartmann-Hahn-multiple quantum coherence and nuclear Overhauser-multiple quantum coherence spectroscopy: application to interleukin 1. beta. *Biochemistry*, **28**:6150–6156, (1989).

[139] E Nogales and S H W Scheres. Cryo-em: a unique tool for the visualization of macromolecular complexity. *Molecular Cell*, **58**:677–689, (2015).

[140] A Bakan, L M Meireles, and I Bahar. ProDy: Protein Dynamics Inferred from Theory and Experiments. *Bioinformatics*, **27**:1575–1577, (2011).

[141] D Bellini and M Z Papiz. Dimerization properties of the RpBphP2 chromophore-binding domain crystallized by homologue-directed mutagenesis. *Acta Crystallographica Section D: Biological Crystallography*, **68**:1058–1066, (2012).

[142] K Sugimoto, M Senda, D Kasai, M Fukuda, E Masai, and T Senda. Molecular mechanism of strict substrate specificity of an extradiol dioxygenase, DesB, derived from Sphingobium sp. SYK-6. *PLoS ONE*, **9**:e92249, (2014).

[143] J S Wischeler, D Sun, N U Sandner, U Linne, A Heine, U Koert, and G Klebe. Stereo- and regioselective azide/alkyne cycloadditions in carbonic anhydrase II via tethering, monitored by crystallography and mass spectrometry. *Chemistry – A European Journal*, **17**:5842–5851, (2011).

[144] S Chakravarty and K K Kannan. Drug-protein interactions: refined structures of three sulfonamide drug complexes of human carbonic anhydrase I enzyme. *Journal of Molecular Biology*, **243**:298–309, (1994).

[145] C-Y Chang, S-Y Lyu, Y-C Liu, N-S Hsu, C-C Wu, C-F Tang, K-H Lin, J-Y Ho, C-J Wu, M-D Tsai, and T-L Li. Biosynthesis of streptolidine involved two unexpected intermediates produced by a dihydroxylase and a cyclase through unusual mechanisms. *Angewandte Chemie International Edition*, **53**:1943–1948, (2014).

[146] R C Lua and A Y Grosberg. Statistics of knots, geometry of conformations, and evolution of proteins. *PLoS Computational Biology*, **2**:e45, (2006).

[147] R Ramakrishnan, J F Pekny, and J M Caruthers. A combinatorial algorithm for effective generation of long maximally compact lattice chains. *The*

Journal of Chemical Physics, **103**:7592–7604, (1995).

[148] R J Wilson. *Introduction to Graph Theory*. Prentice Hall, (2010).

[149] Y Diao, C Ernst, A Montemayor, and U Ziegler. Generating equilateral random polygons in confinement iii. *Journal of Physics A: Mathematical and Theoretical*, **45**:465003, (2012).

[150] E J Rawdon, J C Kern, M Piatek, P Plunkett, A Stasiak, and K C Millett. Effect of knotting on the shape of polymers. *Macromolecules*, **41**:8281–8287, (2008).

[151] J Rudnick and G Gaspari. The asphericity of random walks. *Journal of Physics A: Mathematical and General*, **19**:L191–L193, (1986).

[152] D N Theodorou and U W Suter. Shape of unperturbed linear polymers: polypropylene. *Macromolecules*, **18**:1206–1214, (1985).

[153] V Turaev. Knotoids. *Osaka Journal of Mathematics*, **49**:195–223, (2012).

[154] J Dorier, D Goundaroulis, F Benedetti, and A Stasiak. Knoto-id: a tool to study the entanglement of open protein chains using the concept of knotoids. *Bioinformatics*, **1**:3, (2018).

[155] J Cantarella, H Chapman, and M Mastin. Knot probabilities in random diagrams. *Journal of Physics A: Mathematical and Theoretical*, **49**:405001, (2016).

[156] E Orlandini, M C Tesi, E J J van Rensburg, and S G Whittington. Asymptotics of knotted lattice polygons. *Journal of Physics A: Mathematical and General*, **31**:5953–5967, (1998).